The Theory and Practice of Co-active Search

by Mark Truran, B.A. (Hons.)

Thesis submitted to The University of Nottingham
for the degree of Doctor of Philosophy, 30th June 2005

Contents

List of Figures

List of Tables

Abstract

Retrieval systems implementing query-based search routinely encounter difficulties related to lexical ambiguity. Frequent attempts have been made in the fields of Information Retrieval (IR), Artificial Intelligence (AI) and Natural Language Processing (NLP) to address this problem, yet it remains unsolved. This thesis introduces a novel methodology - known as *co-active search* - that expresses the users of a system as the paramount resource for resolving query term ambiguity. A co-active system resolves the word sense of a given query by recording and analysing navigational behaviour. Captured data is modelled in a multi-dimensional feature space and evaluated through cluster analysis. The resulting sense clusters are then used to auto-categorise sets of search results, enabling the user to quickly select results matching the appropriate word sense for a particular query term. This method addresses lexical ambiguity in a democratic and adaptive fashion, so that a particular word sense can evolve as common usage dictates.

Also introduced is the SENSAI system, a concrete implementation of the co-active search approach. Various components of the system are described, and a full user test documented. Experimental results from this test suggest that the most appropriate way to deal with lexical ambiguity in query-based search is to devolve the sense resolution process to the users of that system - that the business of determining what words mean should be left in the hands of the people using them.

Acknowledgements

I would like to thank my supervisor, Helen Ashman, for comments and suggestions throughout my research. I would also like to extend my appreciation to the members of the Web Technology Research Lab for their contributions in discussions, their shared knowledge, and their participation in various user trials.

Importantly, I *must* mention James Goulding, who worked alongside me tirelessly during the conception and realisation of co-active search, routinely working marvels from his keyboard.

Finally, these acknowledgements would not be complete if I did not thank my Mother, Father and Sister, who have supported me throughout my higher education. It's been a difficult ride at times. Thank you for your steadfastness.

This work was funded by the EPSRC, grant number 20164.

For April, without whom all is lost

Chapter 1

Introduction

In his remarkable prescient article, Vannevar Bush famously described a near-future in which the intellectual capabilities of human beings engaged in information seeking tasks were enlarged by a desk-based system. Known as MEMEX, Bush's machine enabled the storage, retrieval and concordance of disparate sources of information. Levers, slides, buttons and microfiche combined to provide an environment that augmented its user - 'an enlarged intimate supplement to [an individual's] memory' [50]. Despite later work [51], Bush's vision was never implemented in any recognisable form, but his writings delineated a tangible goal for the development of home and office computers later that century. This goal was the development of technology which would facilitate our work by releasing us from the limitations of our own memory.

In certain respects this goal has been achieved. A stand-alone computer system offers its user the opportunity to organise and structure any given body of information for later use. This being said, the storage pattern selected by the user is very much a matter of personal preference, and the ease with which data can be located at some later point may vary. On the one hand, information can be stored in such a way that it accords with the individual's mental model of how such data should be organised, creating a comfortable information space in which to work which facilitates re-use (i.e. appropriately named folders, informational file names etc.). On the other hand, data can be stored in a more chaotic and unstructured manner, conforming to no real coherent strategy, which hinders location and data re-use. However, the problems associated with the location of data on a stand-alone personal computer

are not technology problems *per se*, they are people problems. This is to say, as with more physical environments (e.g. the place you live) the owner of the space dictates the way in which it is organised, and henceforth suffers the penalty where structure is lacking. Current technology, with customisable file search and ubiquitous browsing/preview functions, is more than adequate for all but the most demanding data-finding tasks assuming even a little thought prior to the first file operation.

However, while the retrieval of information from one's own computer is a relatively trivial problem, or at least a problem with a pre-existing solution, the difficulties associated with the location of information stored as part of a computer *network* are formidable. The problem here is undoubtedly one of scale. With connectivity to any large network of computers, the amount of digital information available to any competent user has increased exponentially, with the adversities inherent in the routine search and location of any one particular file growing apace. Gone are the hopes of some machine that could summon any small fragment of information from the ether with 'exceeding speed and flexibility' [50]. This has been replaced with the view of a computer as a tool for laboriously searching the volumes of data produced by a profligate society of users. It is in this respect that Bush's vision has faltered, for if the personal computer is the memory of the individual that owns it then the Internet is the memory of the entire globe. A victim of its own success, this global memory is now too large to be searched with any degree of certainty in the outcome [19].

1.1 Statement of Problem

The search engine is rapidly supplanting traditional office software tools as the mainstay of user productivity. Search engines are now so important to our routine lives that their use is beginning to affect our working vocabulary[1]. Serving users with results chosen from a fluid collection of almost 7 billion documents mere seconds after a query is submitted, the popularity of certain web sites offering hypertext search

[1]Consider the emergent verb *to google* - which means to investigate a particular thing using a search engine.

is testament to the fact that most users visiting these sites come away with their information need satisfied.

However, despite the prodigious commercial and ideological penetration achieved by the humble search engine, there remains one problem with query-based information retrieval engines that has yet to be addressed. This problem is directly related to the way in which we use language and flows from the commonplace fact that one single word may have several meanings. This is the problem of *lexical ambiguity*, which has serious implications for any search engine implementing query-based search and can impact directly on retrieval quality. For example, where a query contains the word *'bat'*, is the user interested in

1. Documents relating to an item of sports equipment (i.e. a cricket bat) *or*
2. Documents concerning nocturnal flying mammals belonging to the order *Chiroptera*.

Thus, the problem in its barest form is this - assuming a query with an ambiguous search term, can we automatically deduce the correct *sense* of that troublesome word, and does this deduction ultimately improve the quality of retrieval effectiveness? This is the problem this thesis aims to address.

1.2 Organisation of this thesis

The structure of this thesis mirrors the slightly unusual process through which the research evolved. It began with a project centred on the development of an automatic hypertext generation system called LIAR (Link Inference and Ruleset engine). The goal of the LIAR project was to create a software application that would complement any hypertext editor (with a suitably flexible API) by serving automatically generated links between any similar web pages in a given collection. The LIAR engine was to be a 'link by example' system, that is to say, a user would be asked to provide the system with some manually-crafted links, and the system would then derive/generalise a set of rules for the creation of new links across the collection [14]. This set of rules

could be added to at any time, and any given rule modified or even annulled by a counter-example.

In anticipation of implementing LIAR as a working system, a full literature review of the relevant field was completed. This review suggested a taxonomy of automatic hypertext generation (AHG) systems that extended to three separate categories. The full text of this literature review can be found in CHAPTER 2. Of particular interest with respect to this thesis as a whole are those sections addressing the difficulties inherent in the resolution of meaning where a particular word is *ambiguous*.

When implementation of the LIAR engine began in earnest, it was very quickly realised that any link generation system would need careful supervision by a human editor. The purpose of this supervision would be to manually screen for errors and correct any linking mistakes as they arose. Consideration of this author-feedback process prompted an examination of the techniques used in information retrieval to gather data from the user concerning the relevance or non-relevance of a given set of search results. With a view to exporting or adapting some of the techniques encountered at this stage of the research directly into the LIAR engine, a further literature review was produced. This review presented a straightforward guide to the key variables that can affect the performance of relevance feedback in relation to an information retrieval system, examining key factors such as user involvement, effectiveness and quality evaluation. The work on this subject can be found in CHAPTER 3.

Now came the change in direction, which was a slow process and moved by increments. Having spent a considerable amount of time considering the problems associated with semantic ambiguity and the the salutary effect of relevance judgments upon retrieval effectiveness, speculation began on a relevance feedback scheme that worked on a larger scale than the usual per-person approaches and specifically addressed the problem of *meaning*. Inevitably, this meant that the LIAR engine was tacitly abandoned, or at the very least, postponed indefinitely. The justification for this decision was very clear at the time, and remains clear to this date. The problems associated with the derivation of meaning are sufficiently grave to condemn any new AHG system almost from its very inception to fundamental and repeated errors related to word sense. To tackle this problem at its root, the emphasis of

the research was therefore re-directed towards the development of a reliable sense resolution system.

Experience with automatic hypertext generation systems and the field of information retrieval in general indicated that the key to such a system did not lay in the statistical observation of word frequency or probabilities, nor the cultivation of expert knowledge bases, and neither could it be found in some paradigm-altering shift in the way in which we share resources. Rather, the tool through which the meaning or meanings of an ambiguous query term could be unlocked lay in the conglomeration of distributed opinion. Another way of stating this observation is to affirm that as in reality, the meaning of one particular word is determined primarily by consensus. CHAPTER 4 outlines the theoretical underpinnings for a system (known as *co-active search*) which is capable of capturing this consensus through the interactions of users with a search system. In CHAPTER 5 the hypothesis that retrieval effectiveness is increased where sense categorisation information is available is put to the test. CHAPTER 6 demonstrates that the application of co-active search techniques is not limited to one particular type of search media (i.e. hypertext documents), but can be applied to a number of other search environments with only minor adjustments. Finally, CHAPTER 7 outlines several possible and promising extensions to the core theory, including its possible application to automatic hypertext generation systems.

1.3 Summary of Contributions of this Thesis

The contributions made by this thesis are as follows:

1. A detailed survey and an applied taxonomy of automatic hypertext generation systems.

2. A substantial review of the key factors affecting the relevance feedback process in information retrieval.

3. An introduction to the theory of co-active search, which can be considered as relevance feedback writ large.

4. A concrete implementation of these theories presented as a working system, with a particular emphasis upon hypertext and image retrieval.

1.4 Final Note

This chapter closes, as it opened, with a quote from Vannevar Bush. Almost fifty years before a practical implementation of his vision was available to the general public, he described in essence one of the fundamental aspects of hypertext, that of the *association*. He said:

> 'The process of tying two items together is the important thing'. [50]

As will be demonstrated in the following chapters, this is as good a summary of co-active search as one could hope to find.

Chapter 2

Automatic Hypertext Generation Systems

2.1 Introduction

The automatic creation of hypertexts has many documented advantages [3]. One key benefit is the soundness of the hyperlinks an automated system can produce. Whereas hyperlinks created by human operators lack much in fundamental rigour and robustness [96, 149], link creation when executed by some software process is typically exact and predictable, providing the algorithm informing the link creation process is demonstrably sound.

This chapter surveys three distinct methodologies employed in the automatic creation of hypertexts. Systems are broadly categorised according to the loci of their analytical focus. This produces a threefold division. First, there are systems which exploit the internal structure of a document (structural systems). Second, there are systems which address the frequency of words inside the document (statistical systems). Third, there are systems which attempt to glean some meaning from the text itself (semantic systems). This division is illustrated in Figure 2.1.

2.1.1 Scope of this chapter

This chapter does not discuss the techniques employed when translating or decomposing large text files into workable hypertext nodes. Although an interesting topic in its own right, it is better documented elsewhere [243, 106]. The primary focus is that of *automatic hyperlink creation*. The systems described herein typically receive as

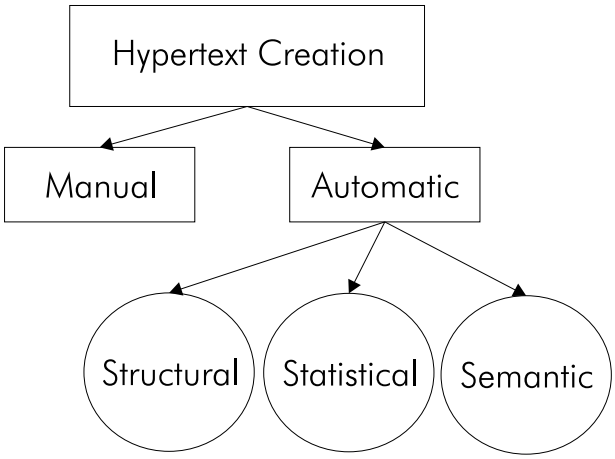

FIGURE 2.1: A tripartite division of AHG systems

their input a large, relatively static document collection. The system then performs some manner of document analysis, and hopefully transmutes this input into some corresponding hypertext of the same material, creating links between documents or parts of documents according to some underlying associative rationale. Links created by an AHG system are therefore *pre-computed links* according to the nomenclature suggested by Ashman [16]. This chapter does not address *dynamically computed links*, that is, those links that remain uninstantiated until some later system event (e.g. traversal by a user). Interested parties are directed to [83, 302, 15] for further discussion.

2.2 Structural Systems

Structural AHGs utilise the internal *logical structure* of documents to generate links. Structural information can be loosely defined as any part of a text file indicating the array of its logical components, for example the disposition and boundaries of sections or chapters. Typically, structural AHG systems are designed to link one recognisable structural element (A) to another (B), although B is occasionally derived by parsing the text contained within A. The link $A \rightarrow B$ can obviously be of either inter-document or intra-document type.

Raymond & Tompa used a structural approach when implementing a 'hypertext-like front-end' for an electronic version of the Oxford English Dictionary [240, 214]. Combing the text files for appropriate link source anchors, the authors selected the notation indicating a cross-reference as 'a substantial source of hypertext links'. However, they were unable to automatically determine a suitable link target by parsing the actual text of the cross-reference. This was, they concluded rather gloomily, a problem requiring 'complex semantic analysis for resolution'.

Glushko documented a similar large-document translation in the paper *Design Issues for Multi-Document Hypertexts* [118]. The task at hand was the integration of a very large engineering encyclopaedia (the *Compendium*), with a 388 page military specification for human engineering design (the *Standard*). Glushko exploited the explicit structure of the documents to create synthesised table of contents, hybrid indices and other 'composite access points'. A rationale for the creation of links was sketched which favoured intra-document links, that is, links which join one part of a document to another part of the same document (for example, footnotes, notations etc.).

> These links are the easiest to identify when converting existing texts to hypertext and they are probably the most usable and useful as well. Since good writers use these structural and rhetorical devices in predictable ways, it follows that a reader of a hypertext will find them predictable as well. [118]

Glushko considered the prospect for identifying and creating links between two *different* documents, in this case the Compendium and the Standard, to be a different matter altogether. He claimed the difficulties that arose when contemplating the creation of *implicit inter-document links* were quite fundamental, and that the task of identifying these links should be left to the hypertext reader. The problem, he declared, was the *subjective* nature of association [95]. One might construct a clever process for identifying inter-document links automatically. The links produced by this process might completely satisfy a well-reasoned set of assessment criteria (relevance, conciseness, coverage etc). However, a subsequent user of the hypertext, with different ideas about how topics 'fit' together, might *still* find the links irrelevant:

> Like a used textbook, some of the highlighting and margin notes may be useful to another student, but may be distracting or misleading at other times. [118]

Importantly, Glushko had recognised that one person's idea of an appropriately linked text might be a 'spaghetti-document' to another user [100]. Therefore, he limited himself to rather safer territory, and emphasised the importance of links within documents.

The construction of inter-document links is occasionally facilitated by some unusual aspect of the hypertext material. When Anderson et al. constructed a prototype for reading the Unix network news through a hypertext browser, they were able to utilise the unique Unix identifiers imprinted on each article [11]. By creating hypertext links between articles and follow-up articles (determined by scrutinising the ID numbers), a hypertext was created that reflected the ancestor-descendant relationships of the raw materials.

Another example of media particularly well suited to automatic conversion into hypertext is that of system or software documentation. Complex or highly technical hardware/software systems are often accompanied by large instruction manuals. Often these manuals have a very definite internal structure designed to improve information retrieval speeds. In addition, they typically feature a self-referential framework of citations (e.g. *for more information about full text parsing, see §6.2*). When considered together, these features describe a type of document ideally suited to automatic processing into some hypertext equivalent.

Franke & Wahl made similar observations when developing a hypertext version of the UNIX `man` pages [106](see also [59]). Stored as *nroff* text files, `man` pages provide a user of the UNIX operating environment with explanations and usage instructions for commands understood by the system. Converting these files into a hypertext entailed translating each file into a number of separate nodes, each identified by the structural clues contained within the text (section, sub-section etc.) These nodes were then processed for links using a simple cross-referential rule. If node A cites a help topic contained in node B, then a new link $A \rightarrow B$ was created.

The authors observed that the quality of the links created using this process

scored a 'high degree of satisfaction' with test users, but admit that some manual intervention and/or editing was required in a supervisory capacity.

> To create hypertext that is easy to read requires some human expertise. A reasonable approximation may be generated automatically but an ideal solution requires associative thinking and subjective decisions. In fact, an optimal solution is impossible to define because the decisions made by the editor are subject to his or her interpretation [106].

In this respect the authors are in accordance with the findings of Rearick, who suggested that systems capable of automatically generating hypertext should work in a *suggestive* capacity [241]. That is to say, they should routinely defer to a human editor charged with final approval.

Further examples of media particularly amenable to automatic conversion to some hypertext equivalent because of their explicit/regular internal structure can routinely be found in the literature - Wilson assembled law cases and secondary materials for the GUIDE hypertext system [38, 316]; Frisse converted the prestigious *Manual of Modern Therapeutics* [221] into a card metaphor searchable hypertext [107]; Nunn et al. used automatic transformation techniques to create an on-line manual for REXX (restructured extended executor), a scripting language for IBM VM and MVS systems [220]; and the WHURLE system utilised structural information to provide auto-navigation for its learners [33].

Providing a valuable synthesis, Furuta et al. [114, 113] advanced a distinct and describable class of 'medium grained information sources' particularly suited to automatic transformation. A membership pre-requisite of this class is a structured document representation [112]:

> The structured document representation identifies the logical components of the document, separating out the specification of the physical placement of the component of the page. The representation is object-based - higher level objects are formed by composing lower level objects.[113]

Members of this class include text documents defined using some generic mark-up languages, database representations of information (which necessarily identify separate logical components), and other documents exhibiting a tightly constrained syntax

(for examples, computer programs, card catalogues). Interestingly, it would appear that paper-based documents *can* belong to this class providing they contain adequate and detectable representational structures (to inform later transformations) [200, 201]. .

One slightly unusual structural AHG system should be mentioned at this juncture. Nentwich et al. developed a rule-based link generation application service known as xlinkit [215]. Given a set of resources and a set of rules, both described in XML (Extensible Mark-up Language) [35], xlinkit verified the consistency of the information contained in the former according to the constraints expressed in the latter. A constraint identifies some relationship that should exist in the data. So, for example, where a hypothetical set of University resources contains one list of modules offered to undergraduate students, and one list of modules offered to postgraduate students, we might wish to enforce a constraint that asserts the module description should be identical in both lists.

The xlinkit service used a 'set-based rule language' to 'express consistency constraints between distributed documents'. XPath [86] expressions were used to select and compare sets of elements from comparable resources. Link generation was dependent upon the results of these comparisons. For each comparison, rather than a Boolean `true|false` statement, xlinkit returned a link:

> We use hyper-links called consistency links to connect consistent or inconsistent elements. If a number of elements form a consistent relationship, they will be connected via a consistent link. If they form an undesirable relationship, with respect to some rule, they are connected via an inconsistent link.[215]

Consistency links were stored in an XLink [87, 88] link-base and provided 'diagnostic information' for the individuals responsible for maintaining the resources. Subsequent actions could include resolving any detected inconsistencies (by editing the resources which are 'pointed at' by *inconsistent* links), or allowing the inconsistency to persist. The consistency links could even be 'folded-in' to the resources using a link-base processor known as *XtooX*, thereby producing 'a web of inconsistency information between files'.

The xlinkit consistency checker clearly differs from the applications previously discussed. The consistency links it manufactures fulfil a different role to the inter-document links produced by Raymond, Glushko, Frisse and others. These authors designed systems that produced navigational links for some equivalent of a hypertext end-user, hoping to make a large corpus of information more accessible. Nentwich et al., on the other hand, produced links of interest chiefly to the actual owners/editors of the hypertext resources, who are perforce interested in any application that shortens routine tasks designed to enforce consistency. However, this difference in link end-use does not affect its system classification.

2.3 Statistical Systems

In the section above it was described how certain structural features of documents in a collection could be identified and used as the source or destination points of a hypertext link. In this section the *statistical* approach to automatic hypertext creation is discussed. This method does not involve analysis of the syntactic constructs present in a given document, but instead revolves around mathematical functions relating to the occurrence of each word (or *term*) in the documents involved. These functions are resolved into some determination of what a particular document is concerned with (sometimes described as *aboutness*), and this determination is subsequently used to inform the construction of hypertext links.

The statistical approach to automatic hypertext creation developed from the field of information retrieval. The classic IR task boils down to this: Given a *user information need*, often expressed as a set of space-separated terms, find a small number of related documents amongst a much larger document collection in a relatively short space of time [18]. A web search engine is one example of an IR system. It converts the text strings that are supplied by the user into a series of suggestions about potentially informative web pages.

Automatic hypertext generation systems that use statistical methods borrow many of the techniques developed for classic IR tasks, but differ slightly in terms of input and output specifications. An IR system accepts a *query*, and returns a list of related

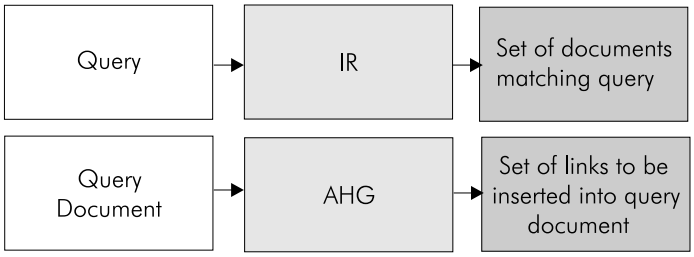

FIGURE 2.2: Comparison of information retrieval and statistical AHG systems

documents from a given collection. A statistical AHG system typically accepts an entire document (which we could label the *query document*), and returns the *same* document with a number of links inserted. These subtle differences in requirements are illustrated above in Figure 2.2:

Having explained the differences and similarities between an IR system and statistical AHG system, it is now time to turn to practical techniques. Luhn provided the theoretical underpinnings of modern information retrieval. He famously observed:

> ...that the frequency of word occurrence in an article furnishes a useful measure of word significance. It is further proposed that the relative position within a sentence of words having given values of significance furnish a useful measurement for determining the significance of sentences [196].

Luhn used this assertion to develop a technique for the automatic creation of abstracts. In his scheme, the term frequency tf for every term in document d was calculated. Analysing the distribution of tf in d, an upper and lower limit was established to exclude those terms that occurred too frequently or very infrequently in the document. These terms were categorised as *non-resolving* function words; that is to say, they were deemed unlikely candidates for inclusion in an abstract because they could not contribute significantly to the article [305].

The remaining terms were *weighted*. Luhn proposed that a term was more likely to be significant if its frequency count occurred roughly equidistant from the upper and lower cut-offs mentioned above. Accordingly, the tf values for these 'ideal' terms were increased at this point. Finally, each sentence was scored for significance (so

that the score of sentence s of length l in terms was equal to $tf^1 + tf^2 + tf^3 ... tf^l$). By selecting the sentences with the highest total score, Luhn was able to effectively abstract the contents of a document with no prior knowledge of its contents.

Adjusting term frequency data with a *weighting scheme* is quite common in information retrieval. One of the simpler weighting schemes normalises against document length [263]. A set of documents of unequal length will produce term frequency distributions that do not fairly represent their respective contents. Therefore, the term frequency tf is usually divided by document length l, so that the weighted score for term x in document d is tf^x/l^d. The precise manner in which the length of each document is actually represented, whether it be in characters, bytes or some other measurable unit, is irrelevant, provided the scheme is applied uniformly across the collection.

Sometimes, a weighting scheme is directly related to the logical structure of the documents being evaluated. For example, Anderson et al. designed a similarity function that weighted term frequency data using node-specific criterion [11] . Each term occurrence was statistically important, but words occurring in certain structurally significant areas of the document (subject field, keyword field, header etc.) were considered more likely to represent document content. Accordingly, these terms were given significantly higher weightings than terms occurring in other parts of the document.

Sometimes the weighting scheme implemented by a particular information retrieval system is related to the underlying hypertext model. In [107], Frisse hypothesised that in the context of a rigidly hierarchical card-metaphor hypertext system, an informative weighting of a card c was composed using two factors:

- The intrinsic value of the terms in c itself.
- The extrinsic value of the cards immediately descendant to c.

This extended weighting scheme was designed to target

> ...parent cards that might not contain any of the query terms, but which might be ideal starting points because of the high fraction of query terms present within their immediate descendants [107].

Obviously, this extended weighting scheme raises significant questions regarding processing costs and response time. To calculate the extrinsic weight of a card c, d further weightings must be figured, where d is the number of descendant nodes of c. Given a hypertext of sufficient depth, a formerly inexpensive and simple weighting operation rapidly becomes quite dear.

Fuller et al. examined these questions when constructing hyperbases of large structured document collections, concluding that the solution could lie in appropriate pre-processing [110]. Each node could contain, alongside its primary media, meta-information such as its *node level* in the structure and a *digest* of its descendant's. This meta-information could then form part of the node-weighting scheme, removing the need for expensive leaf-bound traversals. Although not extensively tested with real users, this scheme was reported as producing 'sensible' results in experimental trials [110].

A more common, general-purpose weighting scheme is known as inverse document frequency [155, 268, 5]. IDF measures the number of occurrences of a word across a document collection. It provides a simple method of diminishing the importance of words that are statistically common, and therefore less useful when attempting the task of document disambiguation. Another way of stating this is that IDF is related to a term's specificity. Where D is the number of documents in the collection, and d is the number of documents that contain a term t, the IDF factor of t in the collection is $log(D/d)$. The higher the IDF value of a term, the more unique that word is across a document collection. The term frequency of a particular term is often multiplied by its inverse document frequency to provide an appropriate composite weighting. This scheme, $tf \times idf$, provides the backbone of many common information retrieval systems [262, 6, 295].

There are three important text operations that are usually carried out prior to the calculating of term frequency. The first is the application of a *stop list*. This is a file of frequently used terms that lend nothing to the actual meaning of the document, such as propositions, conjunctions, pronouns, and so on. This file is applied as a filter, allowing the extraction of common words from a document before further analysis takes place. Stop lists typically contain between 40 to 300 features, and their

application can significantly reduce the size of a document prior to processing [265].

The second measure is known as *stemming*. Prior to the calculation of tf, each term is passed through an algorithm that reduces it to some agreed morphological minimum. A typical stemming operation would, for example, reduce the terms *mountaineer* and *mountaineering* to the same root. Subsequently, these conflated features can be counted as two instances of the same term. Stemming reduces document length and typically improves the quality of the results. There are several published algorithms that can be used [192, 222, 206], and a corresponding list of techniques for evaluating their output [129, 223, 224, 160].

The third operation is known as *expansion*. The purpose of expansion is to maximise the probability of successful retrieval by inflating the query document terms. One common and quick method of expanding a query term t is to search the document collection for t and all synonyms of t [128]. So, for example, if a query document contained the string *car clubs*, the expanded query document might read *automobile car vehicle association club federation*.

Once these text operations are completed and the term frequencies of the query document have been calculated, we can compare that document with every other document in the collection. Rather than calculate tf values for every document in the collection each time a query document is received, most statistical AHG systems generate some kind of *index*. An index (in information retrieval parlance) shares many properties with the paper index we could find in the back of a book. For example, it will feature *value-location mapping*. That is to say, it will provide a list of text strings (terms), and the locations in which those terms can be found. However, rather than display a page number, as in the case of a book index, this location value is more likely to take the form of some internally-understood file location or pointer.

The electronic index will also be *inverted*, which means that it will be internally sorted on the terms that it contains, rather than the locations it points to. This inversion is crucial. Just as inversion speeds the rate at which a human user can look up an entry in a paper index, so it enables a computer to quickly find an entry in an electronic index. Searching through an inverted index, a computer can employ the binary chop algorithm to locate the required entry. This algorithm reduces the

complexity of the search task from n (in the case of a simple search), to $log(n)$ (where n is the number of items to search).

Building an index takes place in advance of the first query submitted to the system, and can be undertaken during periods of low computer demand. Having a pre-built index significantly reduces the response time of an IR/AHG system, and requires little more than a simple application of the statistical measures described above. For a document collection c of size n we build an index in the following way: For 1 to n do:

1. Apply the stop list to a single document d. Stem all features.
2. Derive tf for each unique term.
3. Apply any applicable weighting scheme.
4. Where the tf for any term t exceeds a specified threshold add the pairing (t, d) to the index. Where t is already present in the index, conflate that entry into a one-to-many relationship $(t - d^1, d^2)$.

Determining the threshold value for an index is often a trial and error process requiring hands-on adjustment. Sometimes an arbitrary threshold value is set and then 'fine-tuned' depending on the size and quality of the index produced. In other situations, the threshold value may be determined with reference to the number of index entries, so that the threshold is gradually incremented until the generated index contains a subjectively ideal number of entries.

A simple procedure can now be employed to match the query document with one or more documents from the collection. Significant terms featured in the query document can be looked up in the index, and candidate documents featuring these terms located. Hypertext links can then be created, so that the occurrence of term t in the query document is associated with term t in the candidate document.

This is, of course, a particularly simple scheme. There are various other methods for comparing a collection of documents to a given query. For example, consider the vector space model, backbone of Salton's SMART retrieval system [261] (see

also [238, 183]). In this scheme, the submitted query and a candidate document are represented by an array of term-weight pairs. These arrays are converted to n-dimensional vectors, where n is the number of unique terms present in the collection. Final *query-candidate comparison* is achieved using the cosine measure (that is, the angle between the query vector and the document vector) [260]. The cosine measure is defined as

$$\text{sim}(Q_j, D_i) \;=\; \sum_{k=1}^{t} \omega_{jk} \times \omega_{ik} \tag{2.1}$$

where t expresses all unique terms in the collection, ω_{jk} is the weight of term k in the query, and ω_{ik} represents the weight of k in the candidate document. Both term weights are derived using a $tf \times idf$ scheme normalised against document length. This measure produces a numerical value (usually between 0 and 1) which represents the similarity of the query and the candidate document.

2.4 Specific Examples of Statistical AHG Systems

We now move on to specific examples of statistical AHG systems, as opposed to the generalised theory behind them. We begin with the work of Bernstein and his hypertext apprentice [25]. The apprentice was a 'best-effort' construct. It used computationally 'meagre' textual and statistical analysis to discover links in a text. Its understanding of this text was fundamentally *shallow* - it involved nothing more than the simple manipulation of hash tables to define a similarity measure between the page being viewed and every other page in the hypertext. This measure was then used to derive the twenty most similar candidate nodes and suggest them to the hypertext user as possible or 'plausible' targets of links.

A discussion of the performance of the hypertext apprentice revealed that many of the links suggested by the system were 'strikingly relevant', and that the apprentice was able to construct convincing hypertext 'tours' given an initial subject. On the other hand, the apprentice was unable to discover structural links, and occasionally produced rather 'bizarre' or unforeseeable associations. Bernstein concluded that

the shortcomings of the system were balanced by the negligible impact on system resources (32 msec/page on a 68000-based Macintosh computer) [25].

We now move on to the Microcosm link service [101]. One of the more interesting features of this system was known as a *retrieval-link*:

> Microcosm has a facility that allows a user to batch a set of text files and to index these documents. Once this indexing has been done a block of text may be selected and the action, compute link, may be chosen. The system will very rapidly return a number of other documents within the system that have a similar vocabulary to the selected block, in the order of the best match [123].

Following this procedure outlined above, a user of the Microcosm system could browse a document collection by repeatedly identifying topics for further investigation and then following the computed links. This would allow them to construct a potentially unique route through the hypertext material, a path personal to them and contoured by their interests [50].

James Allan also used statistical methods to automatically create and *type* hypertext links [6, 7]. Citing several studies supporting the view that the utility of a hyperlink is increased by some indication of its function [70, 225, 204], Allan implemented an AHG system (using the vector space model) along with with novel procedures for determining the nature of the relationship between two linked items. Using visualisation techniques also seen in the field of database analysis [82], Allan was able to successfully 'label' each 'automatic link' according to his own taxonomy of possible link types[1]. Despite this notable success Allan freely admitted a category of relationships not susceptible to automated analysis / automated linking techniques - links which could not be located without human intervention. He classified these problem cases as 'manual links':

> Manual links include those which connect documents which describe circumstances under which one document occurred, those which collect the various components of a debate or argument, and those which describe forms of logical implication(caused-by, purpose, warning, and so on). [7]

[1] Automatic links *types* identified were revision, summary, expansion, equivalence, comparison, contrast,tangent and aggregate [6].

Allan concluded that the automatic identification of these 'manual' links was a task better suited to the Natural Language Understanding (NLU) research and Artificial Intelligence research communities (see further §2.5).

Other systems in the literature using statistical methods to automatically create links in a text are multifarious. Kellogg & Subhas used a cluster-based approach to create links for a dynamic digital library [162](see also [56]); Lelu employed a spreading activation technique to construct 'hyper-paths' in a document database representing a city-scape [184]; Tebbutt produced a hypertext version of a very large service manual known as POMS (Program Operations Manual System) for the United States Social Security Administration (SSA) [301]; and Blustein used both the SMART engine and Bellcore's latent semantic indexing system [84] to create hypertext versions of scholarly articles [27].

2.4.1 Evaluation

The evaluation of statistical AHG systems (like their IR counterparts) typically centres upon two measurements - recall and precision. Recall measures how efficient the system is at retrieving documents from the collections. The recall of a system in relation to a query *a* is equal to the *number of documents about a retrieved / number of documents about a in the collection*. The precision of a system in relation to a query *b* is equal to *number of documents about b retrieved / total number of documents retrieved*. Various studies have illustrated that these two metrics exhibit oppositional properties, so that an increase in one invariably leads to a decrease in the other [156, 67, 41]. For the user of a statistical AHG this typically entails a choice between a smaller set of possibly incomplete links and a larger set of links more likely to contain errors.

2.5 Semantic Systems

Systems that rely solely on the statistical analysis of the distribution of terms to inform the creation of links suffer problems directly related to the *open texture* of natural language. Terms have a number of different meanings, resolved quite naturally

by human readers who (unconsciously) refer to the context in which those terms are used. A system that treats each occurrence of a term in exactly the same way, regardless of this underlying *conceptual* sense, inevitably makes mistakes. We call this the *word sense* problem.

For example, consider the purely hypothetical news article below:

> Congress Backs Missile
>
> The United States Navy today confirmed that a test firing of their A/U/RGM-84 *Harpoon* missile has validated improvements to their radar guidance systems. A White House aide informed a member of the associated press that several fiscal initiatives had also kept the improvements within the Congressional budget. This news came as a relief to several high-ranking *doves* that nevertheless used the occasion to campaign for greater transparency on large military contracts.

Assuming this article was one of a number of documents being analysed by a statistical AHG system, the words highlighted in italics indicate terms which *might* be indexed as significant. Any links generated by this system associating these terms with another occurrence of said term may or may not capture the particular word-sense as it is used above. For example, a statistical AHG may link the above document with another document citing *dove-like* leanings in the American cabinet (which would be a correct association); however, it may also link the document to an article describing the illegal hunting of Minke whales using *harpoons* (which would be an incorrect association).

It is these word sense problems that semantic AHG systems address. The basic approach is to instill within the AHG system some understanding of each context (or domain) in which the term could be used. For example, the term *dove* is a member of at least three separate domains. It indicates an anti-war stance (when used in a political context). It signifies a winged animal of the family *Columbidae* (when read within a context describing the natural world or fauna in general). It further connotes the concept of peace, (if used as an abstract idea or symbol).

Having identified the various domains to which a term may or may not belong, the next stage of the process is to define rules that enable each word sense to be distinguished. One excellent case study of this process can be found in *'A News*

Story Categorisation Problem' [140]. The system described in this paper used a modified version of the Carnegie Groups Language Craft product [139] coupled with 'knowledge-based' rules to enable the automatic classification of news stories into 'broad topic categories'. The system attempted a topic categorisation of unseen material in two distinct stages. *Hypothesization* described 'an attempt to pick out all the categories into which the stories might fall on the basis of the words or phrases it (*sic.*) might contain'. *Confirmation* described 'an attempt to find additional evidence in support of a hypothesised topic or to determine whether or not the language that led to the topics being hypothesised was used in a way that misled the system'. Both stages are primarily pattern matching exercises - the difference between them is the 'directness' with which they are carried out:

> Hypothesization always looks for the same words and phrases. Confirmation looks for different words or phrases using specific knowledge-based rules associated with each of the topics which have been hypothesised. [140]

Crucial to this two-stage process is the notion of a *patternset*, which is a descriptive term applied to a collection of patterns. A pattern describes one or more words, and that pattern is *matched* if these words occur in the news article being processed. If enough patterns from one patternset are matched then that patternset is hypothetically matched against an article[2]. One article may be matched against several patternsets at this stage.

The next stage reduces these candidate patternsets to the most likely to apply. This is the stage that filters for false positives, that is, words that may initially satisfy particular patterns but only because they are used in a 'misleading' way. These 'misleading' usages are identified using topic specific rules. For example, an article describing a lengthy prize title fight may satisfy a number of patterns in the *war* patternset, leading to that particular patternset to be hypothesised, but this hypothesis would ultimately be discounted during the confirmation stage because

[2]This is somewhat of a simplification. There is in fact a pattern weighting which influences the hypothesization phase of the categorisation process. Matched patterns can carry both a 'possible' and a 'probable' weight. These weights contribute to a cumulative patternset weighting. If this cumulative weighting satisfies some threshold set arbitrarily by the rule developer, then the patternset is hypothetically matched.

- It mentions no countries, individuals or organisations associated with wars.
- It mentions no wars by name (e.g. Hundred Years War).

One problem discussed by the authors (and other researchers) is the burden of 'tedious' effort required to 'train' the system (see also [53]). Time taken to develop the *rulebase* (patternset, hypothesis and confirmation rules) was estimated at six person months. Subsequent to the creation of the rulebase there was also the continuing problem of updating the rules to reflect changes in a particular domain. For example, a patternset governing *war* would have to be updated in the event of a serious conflagration, and new patterns written describing the combatants, the name of the conflict, *casus belli*, principal individuals and so on.

Hayes & Pepper used the techniques described above to assist in the creation of a system known as IMAD [138]. IMAD, or 'integrated maintenance advisor', was described as a 'conceptually indexed hypertext representation of written documentation along with tightly integrated diagnostic subsystems'. IMAD was intended for use alongside highly complex systems that provide voluminous sets of documentation (relating to use, servicing and trouble-shooting) for the engineers tasked with their servicing and upkeep. An example of such a complex system might be a commercial airliner, or a military grade weapons platform.

IMAD featured an 'expert librarian' function. The 'librarian' automatically constructed hypertext links within a documentation set using the procedures discussed above (in relation to news articles), thereby improving user navigation. This AHG function was augmented by diagnostic/repair subsystems that assisted engineers with specific faults to determine the appropriate remedial action (see further [158]).

Prior to the creation of an IMAD rulebase, the developer annotated all relevant documentation. When the system was used to assist in the documentation of the HAWK missile system [161] a 'standard dictionary of terms' describing all major and minor components, systems and subsystems was created.

> Each component in the component hierarchy is labelled with the words and phrases that can be used to describe it both specifically (spar 7(in the left wing), a ft76-903 fastener) and generically (the spar, the bolt).

> The generic descriptions may be supplemented by information about the type of component (fastener, electrical, electronic, hydraulic subsystem etc).[138]

Using these annotations as reference points a rulebase was created. This enabled 'appropriate references to be found and index entries to be made on a conceptual basis, rather than based on the occurrence of specific key words or phrases'. As with the news classification project, significant human effort was expended upon this annotation of the hypertext material and the creation of applicable patternsets.

Lenat et al. later asserted two major problems with expert systems like IMAD. The first is the 'knowledge-acquisition bottleneck' [187]. This describes the lengthy and iterative process of collecting expert opinion and installing it into some given system, a process usually measured in months and years. The second problem is that of system 'brittleness':

> Carefully selecting just the fragments of relevant knowledge leads to adequate but brittle performance: when confronted by some anticipated situation the program is likely to reach the wrong conclusion - a skin disease diagnosis system is told about a rusty old car, and concludes it has measles.[187]

As an antidote Lenat and others laboured on the creation of a massive knowledge base that embodied common sense known as CYC. The result, after one century of person effort, was a system comprised of around 10^5 general concepts/terms and 10^6 common-sense axioms [185]. These concepts and axioms provide a general body of knowledge that could augment expert system level rules, a 'common sense substrate' that describes the world as we unconsciously accept it. A typical CYC truism might be this - *one object cannot rotate clockwise and anti-clockwise at the same time*:

> Such assertions are unlikely to be published in textbooks, dictionaries, magazines, or encyclopaedias, even those designed for children...These assertions are so fundamental that stating them to another person, aloud or in print, would likely be confusing or insulting.[185]

Multiples uses for the CYC system have been proposed [186]. It could supplement the basic grammar/spelling correction features of a word processor; assist in

the construction of richer user models for adaptive hypertext systems; automatically check for logical inconsistencies in the content of databases: and much, much more. Of course, one further and quite interesting possible application for CYC would be its involvement in the automatic creation of hypertexts, possibly alongside the descendants of some of the expert systems already discussed.

There is a wealth of other systems in the literature that use semantic techniques to automatically create hyperlinks. Green used lexical chains to associate sections of text concerning similar subjects [120](see also [213]); In [278] theasaurus-concept structures were used to augment a semantically based TextTiling algorithm [144]; and Valle et al. developed EIPs (Enterprise Information Portals) that 'understood' meta-data descriptions of corporate resources, improving user navigation with automatically created links [303].

2.6 Conclusion

In this chapter a taxonomy of automatic hypertext generation systems has been introduced. Systems that rely upon the internal *logical structure* of documents to generate links are often quite trivial to implement, but they are dependent upon a consistent and well-considered syntax.

Systems that perform *statistical* analysis of the document collection to generate links evolve errors related to the open texture of natural language. Terms have a number of senses determined by context. Because they have no understanding of context, statistical systems frequently misinterpret *word-sense* leading to unexpected/unpredictable results.

Systems that rely on some *semantic* understanding of natural language to generate links *can* address this word sense problem, but are laborious and expensive to construct. The expert systems that have been produced to date might function well within a limited sphere, but can be 'tripped up' by queries originating outside their area of knowledge. Knowledge bases capable of adequate functioning in an unrestricted information environment are still beyond our grasp, although arguably within our reach.

Chapter 3

Relevance Feedback

3.1 Introduction

Software systems frequently rely upon human intervention to correct or moderate their work. Often this intervention takes the form of a re-purposing or redefining of the stated task in response to the system's best effort so far. We call this user-system interaction *feedback*

When information retrieval (IR) systems gather feedback they are usually concerned with the relevance or non-relevance of retrieved documents. This information is used to guide further searches and improve results. This chapter is a survey addressing several important aspects of the relevance feedback process, with particular attention to the vector space and probabilistic methodologies.

3.2 The Need for Relevance Feedback

Information retrieval systems typically take as their input a series of terms known as a *query*. This query in some way formulates and embodies what the user of the search system would like to retrieve information about. The usual output of an information retrieval system is a small series of pointers to a set of ranked documents extracted from a larger collection. These documents represent the system's response to that information need, each document having been determined by the system as being in some way *relevant* to the query.

User formulation of the query is a problematic process. Firstly, the user may have an incomplete understanding of their information need. Secondly, the user may not understand the correct syntax necessary to frame their query, or the commands necessary to interact with the system [30, 31]. Thirdly, the user may have an incomplete understanding of the document collection that is to be searched. The quality of the query submitted by the user will therefore be extremely uncertain - the query is in fact usually no more than a set of speculative criteria the user hopes will describe the document he or she would like to retrieve.

The practical ramifications of these observations are simple - results generated by an IR system will frequently contain incorrect output [180, 4]. It is possible that the query will be too imprecise, resulting in a surfeit of suggestions. The query could also be conceptually ambiguous, meaning the results are confused by documents that contain one or more of the search terms in the wrong context. The query could also be formulated in error, a problem which is particularly pronounced in the context of boolean information retrieval systems [322, 121].

As an antidote to the problems mentioned above, researchers working with information retrieval systems have incrementally developed strategies designed to improve the quality of the submitted query. One such strategy is referred to as *relevance feedback* or *RF*. It has repeatedly been demonstrated across differing implementations using a wide variety of approaches that relevance feedback can improve the performance of an information retrieval system[1] This increase in system performance is superior to that which can be gained from query improvement techniques based on 'knowledge structures', for example, elaboration of the query using a thesaurus [188, 286, 58].

3.2.1 Aims of this Chapter

This chapter provides a straightforward guide to the key variables that can affect the performance of relevance feedback in relation to an information retrieval system.

[1]Interested readers are directed to [251, 264] for a discussion of RF in relation to the vector space methodology; [78, 126] demonstrate its application to the probabilistic model; and [235, 191] evaluate its use in Boolean retrieval systems.

Readers requiring greater theoretical depth are directed to [257]. Those interested in relevance feedback from a software engineering point of view are referred to [125], which describes the algorithms and data structures involved in the process.

In section 2 we categorise the various types of relevance feedback schemes using a criterion based on user involvement. In sections 3-4 we discuss the basic instruments of relevance feedback in relation to two popular retrieval models (vector space and probabilistic). In sections 5-7 we examine several important aspects determining the effectiveness of RF in these systems, including the role of the user and quality evaluation. Finally, in section 8 we present an overview of the field and some observations regarding the possible future for RF.

3.2.2 Scope of this chapter

This chapter will not examine relevance feedback in the context of boolean retrieval systems [242, 28, 236]. Neither do we discuss some of the newer techniques for relevance feedback, for example those employing connectionist models, latent semantic indexing or genetic algorithms (GAs). Interested readers are directed to the following papers for discussion of these topics [90, 171, 84, 26].

This chapter is concerned primarily with RF in relation to text documents. The use of RF is relation to other forms of searchable media (for example still images, audio clips and video files) is better documented elsewhere [283, 115, 310]

3.3 Types of Relevance Feedback Schemes

Relevance feedback schemes can be loosely classified according to the degree of user involvement in the feedback process. The literature reveals a rich spectrum of participation ranging from full user involvement in a given process to those approaches that require none at all. At one extreme of the scale we have the *manual* feedback approach. Here, the user considers a set of search results served in response to an initial query, manually identifies important characteristics of relevant and non-relevant documents (for example, the presence or absence of an important term), and translates these observations into a modified query. The query is then re-submitted and

a set of hopefully more accurate results served to the user. Note that the business of 'improving' the query in a manual feedback scheme is solely the responsible of the search system user. If the user does not possess the sufficient degree of search skill [116] necessary to isolate/translate the desirable properties of relevant documents, then manual relevance feedback is unsuitable.

An *interactive* relevance feedback scheme places less emphasis on the user, but they still play a central role in the feedback process. In such a scheme, the user identifies relevant and non-relevant documents returned by the initial query. This identification process can take many forms, although check-boxes or click-able hyperlinks (e.g. "More results like this") proximate to the document surrogates are common. The search system then suggests a query modification in response to this feedback. This modification may involve a quantitative change in the number of search terms in the query, also known as *query expansion*. It may also involve *re-weighting* of the original query, whereby information contained in the relevant and non-relevant set of documents is used to increase or decrease the relative importance of a given query term[2]. The user responds to the suggested query modification, either accepting the changes verbatim or possibly overriding the system suggestions. The modified query is then run against the document collection and a second set of results are returned to the user. This process continues until the information need is satisfied or the user abandons the search.

An *automatic* relevance feedback scheme functions precisely as above, but without the final 'negotiation' stage. That is to say, once the user has indicated a number of relevant or irrelevant documents (in whatever fashion) the search system does everything else. The amended query may or may not be visible to the user at any given time.

Finally, *blind* relevance feedback schemes (sometimes known as a pseudo- or naive-relevance feedback schemes), perform feedback without *any* user involvement. The information retrieval system accepts an initial query and a set of ranked documents

[2]Query expansion (QE) and query term re-weighting (QTR) are the basic instruments of any relevance feedback process. Any reference to relevance feedback in this chapter includes both techniques unless it is explicitly stated otherwise.

is retrieved. The top n documents in this set are then automatically added to the relevant set (the assumption being that they are likely to closely correspond with the user's information need), whilst all other documents are relegated to the non-relevant set. The initial query is then modified in accordance with this data, meaning original query terms are re-weighted and new query terms introduced as required. This modified query is then run by the system and the results presented to the user. The process continues, as above, until the search is either successful or abandoned by the user.

The four categories of relevance feedback schemes discussed above will all feature in the following paper, although particular emphasis will be placed upon the automatic and blind feedback approaches. The terminology used above will be maintained throughout for the sake of clarity.

In the next two sections, we introduce introduce basic relevance feedback techniques in relation to two popular retrieval models, to provide a context for further discussion.

3.4 Relevance Feedback in the Vector Space Model

As discussed in §2.3, the vector space model is an enduring and successful paradigm in the field of information retrieval, providing the theoretical underpinnings of countless implementations [267, 238, 262, 183]. Its deserved success lies in its relative simplicity. Each term in the query and each term in every candidate document is initially weighted using an appropriate weighting scheme (for example, $tf \times idf$ normalised against document length [155, 268, 5]). The submitted query and each candidate document are then represented as n-length arrays of term-weight pairs, where n is the sum of unique terms present in the collection following the usual text manipulations (stop listing, stemming etc.) [263, 42]. These arrays are subsequently transformed into n-dimensional vectors, thereby facilitating query-candidate comparison using the following cosine measure:

$$\text{sim}(Q_j, D_i) \;=\; \sum_{k=1}^{t} \omega_{jk} \times \omega_{ik} \tag{3.1}$$

where

t expresses all unique terms in the collection

ω_{jk} is the weight of term k in the query and

ω_{ik} represents the weight of k in the candidate document

This measure produces a numerical value (usually between 0 and 1) which represents the angle between or similarity of the query vector and the document vector. Where $\text{sim}(Q_j, D_i) \;= 1$, the query and the document are identical.

3.4.1 The Rocchio Formula

One of the first widely used relevance feedback algorithms for the vector space model is known as the Rocchio formula [252, 251]. Rocchio designed an automatic relevance feedback system which required a user to classify a small set of documents returned by the original query. The representative vectors for the relevant documents were summed and then normalised by dividing the composite vector by the number of relevant documents. The same process was then carried out with the non-relevant documents. The original query was then modified by *adding* the normalised relevant composite vector and *subtracting* the non-relevant composite vector. This had the effect of re-weighting terms contained in the original query whilst simultaneously expanding that query with terms extracted from relevant documents, thus moving the original query vector away from the centroid of the non-relevant documents and towards the centroid of the relevant documents [43]. Rocchio's formula in full is shown below:

$$Q^{new} = \alpha\; Q^{old} + \beta\; \frac{1}{rel\ docs} \sum_{rel\ \ docs} wt_1 - \gamma\; \frac{1}{known\ nonrel\ docs} \sum_{known\ \ nonrel\ \ docs} wt_2 \tag{3.2}$$

where *rel docs* represents the set of relevant documents, *known nonrel docs* represents the set of known non-relevant documents, and $\alpha - \gamma$ are parameters which determine the relative importance of each formula component. (i.e where the ratio $\alpha.\beta.\gamma$ is set

to 1.4.2, relevant terms are twice as important as non-relevant terms, and four times are important as original query terms.)[3]

Ide later extended the Rocchio formula by removing the vector normalisation calculations and stressing the importance of the non-relevant document set [148](see also [264, 137]). One further modification replaced the element incorporating information from *known* non-relevant documents with

$$... - \gamma \frac{1}{nonrel\ docs} \sum_{nonrel\ docs} wt_2 \tag{3.3}$$

The importance of this modification lies in the potential for over-fitting inherent in the original formulation (see further §3.8.1):

> 'The modified formula makes the assumption that all unseen documents are non-relevant, an assumption which is undoubtedly false, but perhaps better than the assumption that the known non-relevant documents are representative of non-relevant documents in general'. [47]

Buckley et al. continued the trend of experimenting with the Rocchio formula by dropping its non-relevant component altogether, achieving a correspondingly better than average performance on the TIPSTER collection [49, 130].

3.5 Relevance Feedback in the Probabilistic Model

Information retrieval systems using the vector space model constitute a significant proportion of published systems, but there exists a large body of work which addresses IR systems dependent upon a probabilistic methodology [203, 108, 73]. Probabilistic IR systems rank documents by the probability that a document is relevant to a particular query. This is in accordance with the Probability Ranking Principle (PRP) [246, 247, 26], which states that

> '...the overall effectiveness of a retrieval system to its users will be on average the best obtainable on the basis of the data available to the system

[3]Interested readers are directed to [324] for a thoughtful discussion of the significance of the Rocchio formula, in particular the relationship between the values of parameters α and β and overall retrieval effectiveness.

> if the system's response to each request is to rank the documents of the collection for the user in order of decreasing possibility of usefulness to him, where the probability estimates are the best that can be made on the basis of those data'. [72]

Where no information with respect to the relevance (R) of document d_i to query q_j is available (i.e. prior to relevance feedback) some way of estimating the probability of the relevance $P(R|q_j, d_i)$ must be found. This begins with some simplifying assumptions. It is assumed that the relevance of a document d_i to a particular query q_j is independent of other documents in the collection. It is also assumed that terms occurring in a query/document are stochastically independent [247, 71]). Next, the query and each document in the collection are represented as binary term vectors $\vec{x}$ so that

$$\vec{x} = \{x_1, x_2, x_3 ... x_n\} \tag{3.4}$$

where $T = \{t_1, t_2, t_3 ... t_n\}$ is the vocabulary of unique terms in the collection, $x_i = 1$ if $t_i \in D_j$ and $x_i = 0$ if $t_i \ni D_j$.

According to [249], each term is now assigned a weighting so that:

$$w_t = \ log \ \frac{p_t(1 - q_t)}{q_t(1 - p_t)} \tag{3.5}$$

where p_t is the probability of term t occurring in a relevant document and q_t is the probability of term t occurring in a non-relevant document. Since there is no concrete information with which to estimate either p or q at this time, for the first search iteration it can be assumed that all the query terms have an equal probability of occurring in relevant documents [78]. Documents are ranked according to the sum of the term weights they contain which are featured in the query, and the highest ranked documents are returned to the user.

3.5.1 The F4 Formula and variants

As relevance judgements become available to the system, the individual term weightings and the corresponding document rankings can and should be adjusted accord-

ingly. In [249], relevance information was factored into a term weighting scheme latterly known as the F4 formula:

$$w_{ij} = log \frac{\frac{r}{R-r}}{\frac{n-r}{N-n-R+r}} \tag{3.6}$$

which is equivalent to

$$W_{ij} = log \frac{r(N-n-R+r)}{(n-r)(R-r)} \tag{3.7}$$

where

W_{ij} is the weight if term i in query j.

r is the number of relevant documents for query j containing term i.

R is the number of relevant documents for query j.

n is the number of documents in the collection containing term i.

N is the number of documents in the collection.

Sparck-Jones later extended this formula to allow for query terms that did not occur in the set of relevant documents [287]. A value of 0.5 was added to each of the four elements in the F4 formula to prevent the calculation of infinite weights. Further refinements to the F4 formula were suggested in [134, 248].

3.5.2 Negative Re-weighting

One the problems attached to the probabilistic methodology is the phenomenon of negative re-weighting. As Harman correctly observed, negative re-weighting leading to mediocre system performance typically occurs when experimenting with relatively short queries and relatively short documents [126]. This can frequently lead to functional difficulties directly related to sampling errors:

> 'This can be a problem for certain queries where a high-frequency term appears in the query, but not in the initial set of relevant documents. If this term is re-weighted, then it is likely to be negatively weighted, and this will cause any later relevant documents containing this term to be badly ranked...this phenomena could explain the uneven performance of the probabilistic model across collections'. [126]

One possible solution has been put forward by Bookstein, who advocated an extrapolated or speculative stance in the early stages of the feedback process [29]. This modification was tested by Harman [126] using the F4 formula and the Cranfield collection of 1400 documents and 255 queries [65, 68, 66]. The modification had little effect on system performance during the experiment, but it seems likely the results were collection-specific rather than general. Cranfield features relatively long queries and documents, and therefore is not an ideal testing ground for a strategy designed to tackle the phenomenon of negative weighting. There are also outstanding concerns with regard to the Cranfield relevance judgements which may have had some small impact on the results [307, 135, 300].

3.5.3 Query Expansion

Thus far, relevance feedback for probabilistic IR systems has been explained solely in terms of term re-weighting. This can be contrasted with relevance feedback using the Rocchio formula, which of course not only re-weights query terms but also expands the initial query with new search terms. Various attempts have been made to add query expansion to the probabilistic model [94]. One such technique involves expanding the query using terms extracted from a maximum spanning tree (MST) [134, 13]. Another general strategy collects non-common terms from relevant retrieved documents, sorted using a variety of measures [133, 128]. However, as the literature clearly illustrates, it has proved unusually challenging to establish a strategy that survives contact with a high number of differing collections. Problems are particularly acute when collections typified by short documents are involved [284].

One further strategy which should be mentioned at this time is the model-based feedback approach suggested by Zhai & Lafferty [327]. Developed from the language modelling (LM) approach to information retrieval advanced by Ponte [229, 230], the authors begin by suggesting that 'expansion-based' relevance feedback techniques using LM theory (such as [208]) create an 'inconsistent interpretation of the original and expanded query' in variance with the LM model. In contrast, they introduce a 'model-based' feedback approach they assert as more in keeping with classical probabilistic

theory [249], designed as part of the KL-divergence retrieval framework [178]. Their experiment documents the testing of both a *generative* model and a *divergence/risk* model:

> 'The first method assumes the feedback documents are generated by a mixture model in which one component is the query topic model and the other the collection language model.... The second method uses ... the query model that has the smallest average KL-divergence from the smoothed empirical word distribution of the feedback documents'. [327]

Evaluation of both methods was positive, with both the generative and the divergence/risk model outperforming Rocchio w.r.t precision on three large collections (AP 88-89, TREC8 (ad hoc) and TREC8 (web)) given 10 feedback documents. This performance was mirrored at 50 feedback documents, although not reported in depth.

3.5.4 Further Reading

Interested readers are directed to [288, 289] for a comprehensive account of the theoretical underpinnings, design and testing of a probabilistic model of retrieval. Concerned with the 'city' probabilistic model proposed in [249], this paper details experiments using several collections (Cranfield, UKCIS [74], NPL [157] and an enlarged TREC collection designated 'T741000X'), unifying a significant body of previously published material and data. Also of note in this respect is Lancaster's large scale investigation, published in [180].

The following three sections examine important aspects of the feedback process. The first section discusses the operational choices surrounding expansion of the query. The second section debates the role and relative importance of a user of the search system in the feedback process. The third section considers factors best considered as *general*, ranging from performance evaluation of a feedback process to common RF maladies.

3.6 Factors related to Query Expansion

Query expansion is a part of the relevance feedback process and can be applied to both vector space and probabilistic retrieval models (see §4.3). The questions raised below are of paramount importance during the implementation of any blind feedback scheme.

3.6.1 Degree of Expansion

The degree of expansion expresses the number of supplemental terms added to the original query by the IR system related to the number of terms in the original query. In certain circumstances, where the query is of a particularly low initial quality, the most appropriate degree of expansion is nil[4]. Lundquist et al. identified a relationship between the percentage improvement in exact precision seen during the query expansion process and the *nidf* of the terms in the initial query (where the *nidf* of a term t is the weight of t in the collection) [197]. This relationship (a correlation coefficient of +0.24) suggested that queries below a certain minimum *nidf* weighting threshold should be completely excluded from the expansion process. This hypothesis was tested using the short versions of the TREC5 queries and discs 2 and 4 of the TIPSTER collection, with a small positive gain.

Another important factor when determining the most appropriate degree of expansion for a particular query is *expansion cost*, being the extended processing requirements of the augmented query, longer queries necessarily equating to lengthier query-document comparisons. This is not a particularly grave problem when the IR task concerned is the off-line filtering of information for a client (i.e a long-term routing task). However, it has significant quality of service (QoS) issues when the task concerned is an ad-hoc search performed by an expectant user. In this case the degree

[4]The problem of when to apply relevance feedback (both QE and QTR) is the special interest of the TREC Robust Track, which looks to 'improve the consistency of retrieval technology by focusing upon poorly performing topics' [309]. Established in 2003, this track investigates the performance of systems set a traditional ad hoc task, with the problem topics for that run drawn from previous TRECs. The Robust track has seen several interesting experiments examining the relationship between relevance feedback and these low score topics [141, 159, 176, 319], and will continue studying individual topic effectiveness in 2005.

of expansion and thus the effectiveness of the query must be finely balanced against the speed of the transaction [47].

Irrespective of processing cost, there is also an expansion *cut-off point* to consider, a boundary beyond which the addition of further expansion terms to the original query results in a deterioration in system effectiveness. There is some disagreement in the literature over the likely coordinates of this cut-off point. Harman described an experiment using a probabilistic information retrieval system [126]. An initial query was run on the system and a set of results returned. A selection of non-common terms were selected from the top 10 retrieved documents. This collection of terms was then ordered by frequency and noise, where noise equates roughly to idf. Query expansion was then initiated with 10, 20, 30 and 40 terms respectively, with both original and expansion terms being re-weighted. After plotting the number of added terms against the percentage change in precision, she found that the ideal number of expansion terms occurred somewhere between 20-40 terms.

By contrast, several papers originating in Cornell suggests that an information retrieval system based on the vector space model can support a much greater degree of expansion whilst still exhibiting significant increases in retrieval effectiveness [43, 46, 48]. Experimenting with queries 101-150 of the TREC Collection and the third subset of the TREC document collection, Buckley & Salton determined the cut-off point for query expansion effectiveness occurred somewhere between 210 and 330 terms [45]. In a further publication, Buckley et al. advanced the hypothesis that there existed a linear relationship between the log of the number of expansion terms and system performance (measured as recall-precision effectiveness) [47] . The results of the experiment, which used several TREC collection sets, the *ltc/lnc* weighting scheme [263], and a slightly modified Rocchio formula with parameters of 8.16.4, certainly seemed to support these claims, with effectiveness clearly increasing with query expansion up to 500 terms.

The difference between Harman and Buckley cannot be explained purely in terms of underlying methodology. A later study (using the vector space approach, disc 2 of the TIPSTER collection and 50 TREC4 queries) showed that system efficiency measured as precision and recall were at their peak when between 10-20 expansion

terms were added [197].

Recent experiments organised by NIST as part of the NRRC Reliable Information Access (RIA) Workshop have emphasised the importance of selecting the correct number of query expansion terms [132, 198]. The workshop united seven well-known retrieval systems of differing methodologies and set them to a common task. These systems conducted numerous runs on data taken from TREC 6, 7 and 8 (topics 300-450) as part of a comparative failure analysis. Also conducted were several informative blind feedback experiments, where 6 groups used their systems to examine the effect of varying the number of query expansion terms upon retrieval effectiveness. Each group conducted 37 runs using the top 20 documents to source the expansion terms and with the number of expansion terms varying between 0 and 100. Preliminary results suggested that retrieval effectiveness could be improved by up to 30% (as against a fixed number of expansion terms) where the correct number of expansion terms (based on the results) were added.

3.6.2 Optimal Number of Relevant Documents

The optimal number of relevant documents from which to extract feedback terms is a further source of contention. As mentioned above, a linear relationship between the log of the number of terms added to a query and the recall-precision effectiveness was demonstrated in [47]. Moreover, in the same paper, the authors suggested a similar relationship between the log of the number of known relevant documents and recall-precision effectiveness. In the experiment, N documents were retrieved for each query, where N varied from 5 documents to 5000 documents. Expansion terms were gathered by selecting the terms that occurred in the largest number of relevant documents (where relevance was determined using the TREC JQ_2D_{12} relevance judgements). The results demonstrated that as the set of known documents N increased there was a corresponding increase in average recall-precision. While this increase in performance was not as marked as the gain in system performance recorded when the number of expansion terms was steadily advanced, it was still statistically significant. The authors suggested that the implications of this mathematical relationship (and the

relationship between expansion terms and effectiveness) should be pursued further:

> 'The fact that any mathematical relationship at all exists between added information and effectiveness opens up an entire line of future experiments. It may be possible to isolate the effects of particular types of information, and explore weighting phenomena in much more detail than the large scale investigations of the past'. [47]

Existence of the linear relationship discussed above was disputed in [197], which suggested a negative correlation coefficient between the strength of the linear relationship between exact precision and the number of documents used for relevance feedback. This negative correlation coefficient implied that as the number of documents used to provide expansion terms increased, exact precision (averaged over 50 queries) would begin to drop. The authors went on to show that the greatest increases in precision and recall were demonstrated when between 5-20 of the top ranked documents retrieved from the TIPSTER collection [130] were used to source the expansion terms used in the experiments. This finding held true when the quantity of expansion terms (selected using an idf term weight method) was doubled.

The RIA Workshop has already been discussed in relation to the problem of determining the correct number of expansion terms for blind feedback schemes (see §5.1). The workshop also occasioned the *bfnumdocs* experiment suite, which saw six of the participating groups conduct 36 runs against the data with 20 terms taken from a varying number (0-100) of top documents [312]. Results published in [211] confirmed two central hypotheses, namely

1. All the systems would demonstrate a tradeoff in the choice about how about how many documents should be used for feedback
2. The optimal number of documents used for feedback would vary from system to system

Three further hypotheses, including the the interesting proposition that query length and the optimal number of documents used for feedback would be negatively correlated, were not supported:

> These results show that short queries or queries with small numbers of relevant documents may need the same number of documents for feedback as long queries or queries with large numbers of relevant documents [211].

3.6.3 Expansion Term Selection

Assuming both the number of expansion terms and the set of relevant documents from which those terms are to be extracted has been agreed, there remains the problem of quality term selection. The difficulty here is to design a process whereby terms which will benefit the query are added, whilst terms which will degrade or defocus it are not (see further §3.8.1). There are, of course, several approaches[5]. Attar & Fraenkel applied clustering techniques to the top ranked documents returned by a particular query, selecting expansion term candidates from the resulting term clusters [17](see also [286, 193, 317, 44]). As mentioned above, Buckley et al. selected terms which occurred in the highest number of known relevant documents, with ties resolved by average document weight [47]. Fitzpatrick & Dent selected expansion terms not from a set of retrieved documents, but a log of previous queries submitted by users [98]. Robertson ordered candidate expansion terms using an analysis of the distribution of each candidate term across both the collection and the known relevant subset [248]. A related technique, using the Kullback-Liebler distance between the two term distributions, was put forward in [57]

Harman tested several term selection algorithms, sorting terms found in the top 10 retrieved documents according to criteria such as *noise* (an idf-like measure), *postings* (number of documents featuring the term) and *frequency* (derived using the total frequency of the term within the collection) [128]. She later returned to the same topic to experiment with probabilistic sorting techniques, as opposed to the statistical algorithms explored in her earlier paper [126]. Probabilistic sorting techniques arrange term order not on the basis of some verifiable occurrence measure (such as tf or idf) but according to the probability that the term may occur in a relevant document. In this vein, Harman examined the 'relatedness measure' devised by Doszkocs [89], a sorting procedure taken from the Muscat system [231], and a scheme taken from Robertson [248]. Her results demonstrated the importance of 'the total frequency factor' - that is, the importance of the total frequency of a term within

[5]Some of them are discussed here, but interested readers are also directed to [258, 304, 134, 151, 244, 93]

the data collection (as opposed to the frequency factor of the term within the set of retrieved relevant documents).

Feedback term selection techniques varying in nature frequently produce equivalent results [58]. However, performance measured on a query-by-query basis can differ significantly:

> 'Retrieval feedback methods employing distinct term scoring functions produce different expanded representations of each query, with large variations on retrieval performance, even when the same methods present a comparable retrieval performance over the whole set of queries'. [58]

Carpineto et al. suggested a combined query-selection method constructed by analogy with ensembling classifiers to attack this problem [58, 181].

3.6.4 Phrases

Relevance feedback schemes often expand the original query with significant phrases as well as meaningful terms. Phrases are initially identified using term co-occurrency data (see further [304]). Once a stop list has been applied across the collection [259, 167], any adjacent non-stopwords form a phrase possibility [46]. These possibilities are subsequently treated as a phrase proper if they co-occur sufficiently often across the document collection (that is to say, if they exceed what is sometimes known as the *minimal document frequency* for phrases determined for that particular collection). For example, Buckley et al. arrived at a final list of phrases by selecting those term pairs occurring in more than 25 candidate documents [43].

Once the presence and frequency of phrases in a collection has been determined (*phrase indexing*), this data can be used alongside term frequency data to effect a query/document comparison. The precise weighting attached to phrases at this point in the computation will of course vary from system to system. One example, taken from [20, 34], determines the Retrieval Status Value (RSV) [219] for a query q and a document d_j in the following way:

$$RSV(q, d_j) = \sum_{\varphi_i \in \Phi(q,d_j)} a_{ij} \cdot b_i \tag{3.8}$$

where

φ is the set of one- and two-word features (i.e. phrases) occurring in query q and document d_j

a_{ij} is the weight of the feature in the document and

b_i is the weight of the feature in the query.

with basic weights determined using the *Lnu.ltn* weighting scheme proposed by Singhal et al. in [282] (see also [263, 271]). For the purpose of normalisation the contribution of phrases to document length was ignored [43].

Once a user provides the system with relevance feedback, both query term- *and* query phrase expansion can be initiated. The number of phrases to add to the original query, given a selection of phrases to choose from, is as much a matter of speculation as the ideal number of expansion terms. Buckley et al. submitted a TREC2 routing run `crnlR1` which expanded a query with 300 terms and 50 phrases, with the importance of the phrases in relation to query terms set at 0.5, Rocchio parameters of 8.16.4, and the selection of expansion terms and phrases determined by their frequency within the known 'relevant' documents [43]. Run `ccrnlR1` performed ahead of other TREC2 routing runs, and the authors extended the experiment to determine an optimal expansion for each query. This involved examining the 158 TREC2 routing runs query by query, and determining retrospectively an ideal set of parameters for each one. Their findings indicated that there exists a troublesome:

> '... distinction between those queries where phrases are useful and those where phrases appear useless: 1 query worked best adding 100 phrases, 6 with 50 added, 2 with 10, 16 using the original phrases only, and 25 using no phrases at all'. [43]

3.7 Factors related to the User

Blind feedback schemes specifically exclude the user of a search system from the relevance feedback process. We now discuss the merits of such an exclusion, and in particular, the significance of additional contextual information qualifying a given query.

3.7.1 User Interaction: The Case Against

Relevance feedback schemes that require user participation share one major drawback. It has been observed, both in relation to relevance feedback and information retrieval as a whole, that users can prove intransigent when required to provide additional information beyond the bare minimum. In fact, what emerges from a good deal of the literature is a picture of information retrieval system use that is characterised by short queries and limited interaction coupled with unreasonably high expectations [77, 254, 168].

> 'Real people in the real world, doing real information seeking and in a hurry, use web search engines and give 2-word queries to be run against billions of web pages. We expect, and get, sub-second response time and we complain when there is no relevant web pages in the top 10 presented to us'. [39]

According to the above, it would be logical (and indeed correct) to assume that the overwhelming majority of published feedback schemes are blind in their operation, simply because of the difficulty in encouraging user cooperation. Even when user co-operation is forthcoming, several studies have shown that human searchers expanding a query in an interactive fashion are often outperformed by blind techniques, typically because they make sub-optimal expansion decisions [21, 199].

3.7.2 User Interaction: The Case For

Speaking for the minority, Koenemann conducted a series of experiments which evaluated the ability of four groups of users without formal IR training to design an effective query for an information filtering (or routing) task [169]. All four of the groups used the INQUERY retrieval engine [54]. All four groups had the same retrieval/routing task, namely the selection of relevant documents from a subset of a TREC2 test collection [131, 127] given two search topics (*automobile recalls* and *tobacco advertising and the young*). The key difference between each group was the availability of relevance feedback and the extent to which it could be manipulated by the user, ranging from a system offering no relevance feedback at all (the control

group), to a 'penetrable' relevance feedback interface that supported user interaction (i.e. the ability to edit system selected expansion terms). He noted that

> '...availability and use of relevance feedback increased retrieval effectiveness; and increased opportunity for user interaction with and control of relevance feedback made the interactions more efficient and usable while maintaining or increasing effectiveness'. [169]

In a later report, Koenemann subsequently confirmed user-oriented relevance feedback as a 'beneficial mechanism', both in terms of system performance and human-computer interaction (HCI) considerations [170]. However, it is worth noting that both cited publications concern the design of a query intended for a routing task (i.e. a persistent information need). Whereas a user might be prepared to spend time and effort providing feedback for a routing task (effort amortised by the flow of requested information over weeks or even months) this is not necessarily the case with ad hoc queries

On a final note, Ruthven [256] conducted his own set of interactive vs. automatic expansion tests which largely concurred with the findings published in [199]. However, the author attributed these findings to a specific lack of support for the searchers, concluding that

> '.. simple term presentation interfaces are not sufficient in providing sufficient support and context to allow good query expansion decisions. Interfaces must support the identification of relationships between relevant material and suggested expansion terms and should support the development of good expansion strategies by the searcher'.

Interactive query expansion might therefore outperform the automatic equivalent where sufficient attention is paid to the infrastructure supporting their search task.

In the HARD track of 2003, it was reported that user involvement in a cluster selection (query expansion) process was indicated when the ranked list of documents in consideration was predominantly relevant, but counter-indicated where the results list contained more than 50% 'noise' [142]. Of course, this left a rather delicate task for later consideration:

> 'Designing a mechanism to automatically identify when to ask the user, and when not to, will be one of the foci of our further exploration of the HARD retrieval model'.

3.7.3 The 'Average' User

Recent studies have hinted that improved retrieval effectiveness may lie in an analysis of external information describing the user, or meta-information provided by the user qualifying the query. The High Accuracy Retrieval from Documents (HARD) track at the annual TREC conference is a response to an observed 'plateau' in the effectiveness of ad-hoc search systems, whereby the year-on-year improvement in typical system effectiveness has bottomed out. This hiatus in system improvement may be related to the generalised nature of the TREC ad-hoc tasks. TREC search tasks have always been pitched at an anonymous 'average user'. This targeting has prevented the use of techniques that identify and serve a particular *type* of user. The HARD track is encouraging just this sort of personalised system:

> 'The goal of this track is to bring the user out of hiding, making him or her an integral part of both the search process and the evaluation. Systems do not have just a query to chew on, but also have as much information as possible about the person making the request, ranging from biographical data, through information seeking context, to expected type of result'. [8]

The HARD track had three stages. In the first stage, participating groups were required to perform an ad-hoc run against standard format TREC topics. In the second stage, groups were able to acquire additional information about the user and the query, contextual hints, biological information, clarifications of ambiguous terms, and so on. The groups then completed a second run against the TREC topics from the first phase, utilising this contextualised information. Results so far have been encouraging. In the HARD 2003 track, Shanahan et al. exploited user feedback to innovate a blind feedback technique based on clusters of documents, demonstrating a 20% improvement for mean average precision (MIP) over the baseline run [275]. In the following year, Sun et al. demonstrated that information from clarification forms could be used in tandem with query metadata to improve search functions [296].

However, the track has not been without its share of negative results [250, 163], with some participants seemingly overwhelmed by the additional information suddenly available to them [23].

3.8 General Factors

In this section we discuss general topics relating to relevance feedback. These topics include two major pitfalls any new algorithm must avoid (known as query de-focusing and overfitting), the factorisation of time in the feedback process, and the performance metrics available to assess system improvement.

3.8.1 De-focusing

Query re-weighting and term expansion are not without risk. As expansion terms are added to the original query and terms are progressively re-weighted the query can become *de-focused*. This means the user's formulation of their information need has become watered down by the feedback algorithm, leading to unexpected or unwanted results. This is a particular danger for systems implementing blind feedback schemes. In this case the likelihood of de-focusing the query is related to the initial quality of the top-ranked documents. If the top ranked documents are by and large relevant, then the resulting feedback will hopefully improve the quality of the query. If, however, the top ranked documents contain an unusually high amount of non-relevant documents, it is likely that feedback will degrade the query. Several strategies involving document or term clustering have been suggested as prophylactic counter-measures against query drift (see §3.6.3). One alternative involves re-ranking the top-ranked documents prior to expansion term selection using automatically generated fuzzy boolean filters [210].

Another strategy, known as *local context analysis*, was proposed in [318]. It rests upon the assumption that 'a common term from the top ranked relevant documents will tend to co-occur with all query terms within the top-ranked documents'. This hypothesis suggested the following procedure for expansion term selection. Assuming a query $Q = t_1, t_2, t_3 ... t_n$, and a set of top ranked documents $t = d_1, d_2, d_3 ... d_n$,

the quality of a possible expansion term (or *concept*) c can be approximated using c's co-occurrence with the terms in Q across the document set t. To allow for the possibility of random co-occurrence, the authors incorporated the frequency of c in the whole collection into the concurrence metric (an unusual formulation of *idf* suggested in [177]). The metric was also normalised over the number of documents in t. All possible expansion features are ranked using this co-occurrence measure, and the best features added to Q.

The authors tested their hypothesis using four differing test collections (TREC3, TREC4, TREC5 and WEST). Local context analysis was shown to outperform the local feedback techniques[6] discussed in the Cornell TREC4 entry [49]:

> 'As we mentioned before, a problem with local feedback is its inconsistency. It can improve some queries and seriously hurt others. A query-by-query analysis on TREC4 shows that local context analysis is better than local feedback in this aspect. Although both techniques significantly improve retrieval effectiveness on TREC4, local context analysis improves more queries and hurts fewer than local feedback'. [318]

Local context analysis was also surprisingly resilient to changes in the number of relevant documents used during the feedback process, in direct contrast to several of the procedures cited in §3.6.2.

3.8.2 Over-fitting

Over-fitting describes the phenomenon whereby a particular IR system modifies the original query too severely, molding it to describe the known relevant documents so closely that its performance on unseen material is degraded. Buckley & Salton cited [323] as a practical example of an over-fitting system, referring to

> '...the need to optimise weights to describe relevance as opposed to optimising weights to describe the known relevant documents. If the latter

[6] Xu & Croft drew an interesting distinction between expensive query expansion techniques which rely on corpus-wide statistics (referred to as *global* techniques) and lighter weight strategies which rely upon more limited data (referred to as *local* techniques)[318]. The following well-known strategies were identified as global techniques for query expansion: Document clustering [286], latent semantic indexing [84] and the construction of similarity thesauri [233]. Examples given of local feedback techniques included [78, 311, 137].

> is the goal, then it is easy to achieve near-perfect retrospective retrieval effectiveness. But this retrospective effectiveness does not translate into effectiveness in retrieving new documents'. [45].

The authors subsequently suggested a policy of 'restrained' alterations, whereby the degree of query change is specifically limited, to prevent over-fitting .

3.8.3 Number of search-feedback cycles

In an *ideal* search system which implements interactive or automatic relevance feedback, as each search-feedback cycle is completed the quality of the original query should increase, with a corresponding gain in system performance. This process hypothetically continues until the user's information need is satisfied by the system - each search iteration drawing the locus of the search progressively closer to its goal[7].

However, a practical IR system running *ad hoc* queries for a live user faces a very real time constraint. Users may become disillusioned or distracted if primary or secondary results are not adequate. Even assuming the initial results are encouraging, there is still a finite number of iterations a user is prepared to commit to. This, of course, begs the question -'How many iterations are necessary for an effective search?'. Harman reported that searching should continue until no more relevant documents are retrieved - a *dead-end* approach [126]. This was coupled with the recommendation that

> '... users should be encouraged to look through more documents, at least another screenful, even if no relevant documents are found, unless they have a clear idea of a better query'. [126].

3.8.4 Feedback and Time

It has been demonstrated that a user's information need will frequently mature during, and in direct response to, the process of searching [179, 150].

> 'Searchers normally start out with an unrefined or vague information need which becomes more sharply focused as their search continues and exposure to information changes their information need'. [39]

[7] We are reminded of Zeno's paradox concerning the impossibility of motion [255].

An information retrieval system which implements automatic or interactive feedback can model this evolving information requirement by weighting relevance judgements against time, so that the judgement with the greatest importance is the latest and the judgement with the least importance is the earliest. This scheme is known as ostensive relevance feedback (O-RF) [55].

As the weighting of relevance judgements deteriorates over time, so should the relative importance of the original query. This observation was verified in [45]:

> '... the relevance evidence from the learning set of documents should be sufficient to determine weights for a query, independently of the original query (assuming enough relevant documents have been found)...If most of the highly-weighted terms are good quality, then the emphasis on the original terms should be decreased to allow the full effect of the learning information to come through'. [45]

Thus, the given weights of the original query terms can wane after the initial search-feedback iteration without a loss of system efficiency *provided* the relevance algorithm provides high quality terms/weights to supplement them. The authors successfully used Rocchio parameters of 2.64.8, meaning that the weighting of terms from relevant documents exceed the weighting of the original query terms by a factor of 32.

3.8.5 Performance Metrics

The benefits of relevance feedback are traditionally scored in terms of recall, precision, or some other published metric which approximates system effectiveness [245, 237]. However, Efthimiadis has suggested that such evaluations are incomplete because they fail to take account of the 'ultimate judges of the performance of the system', the end users [94]. He conducted an experiment in which six algorithms commonly used for query expansion were scored against a set of user-selected expansion terms. His results illustrated a significant difference between the terms preferred by the evaluated users and the terms selected by the best performing algorithms, (van Risjbergen's expected mutual information measure (EMIM) [304] and the $w_t(p_t - q_t)$ formula [248]). A structured analysis of the performance of both of these algorithms revealed that each

had achieved a 68 per cent concentration of user preferred terms in the top half of all selected terms:

> 'Such concentration levels are acceptable for interactive query expansion because the user can browse the list and can recognise terms. However, for automatic query expansion such level might not be acceptable'. [94]

3.8.6 Relevance Feedback and Hypertext Documents

Blind feedback techniques encounter specific problems when the searched collection is composed of hypertext or hypermedia documents. Such documents often contain a significant amount of non-content bearing terms devoted to specifying particular presentation/rendering styles, navigational data, meta-information (such as ownership) and methods or prompts inviting user interaction. Furthermore, a typical web page frequently broaches a selection of topics, with divisions between topics often inferred by some later user rather than positively indicated in a recognised machine readable fashion. The two factors mentioned above make web pages unusually troublesome candidates for blind feedback schemes. A simple blind feedback scheme, which modifies the initial user query by re-weighting the original terms or adding new terms taken from the top n web documents returned, is likely to degrade retrieval results by incorporating non-content bearing or off-topic terms.

[325] proposed a page segmentation technique intended to reduce the 'noise' found in a web document (on this topic, see further [143, 228]). The VIPS (Vision-based Page Segmentation) system combined DOM [12] page analysis with the detection of visual clues (such as font size, page position and colour) to isolate regions of high-quality content within the page. These regions offered significantly better term correlation than the page taken as a whole. A blind feedback experiment was conducted using a candidate segment set extracted from the top 80 web documents returned by a query. Results confirmed that the VIPS algorithm performed favourably against the traditional full-document approach and other DOM-based segmentation techniques [216].

3.8.7 Evaluation incorporating QoS

Rolker & Kramer took a novel approach to user feedback [254]. They accepted the assumption that a typical user of an IR system will often fail to locate an appropriate information source, and will therefore quit the system unsatisfied. This fact they ascribed to two factors - 'too high expectations and wrong use of the system'. The authors suggested that this situation could be improved with the addition of tailored quality of service (QoS) architecture. By encouraging the user to negotiate a retrieval success rate with the IR system in advance of the initial query, a user could be given a more realistic idea of how many searches it might take to reach their information goal. This negotiated figure of search iterations could then used by the QoS Management party, which monitors the service provided by the IR system.

> 'If the service is suddenly disturbed and the negotiated QoS values will not be achieved, the QoS Management tries to adapt the service by reconfiguring it. If there is no way to meet the negotiated QoS values, the QoS Management invokes an action to inform the client'. [254]

Importantly, the authors proposed what Croft [76] might have labelled an adaptive document retrieval system. Relevance feedback to their system changes its constitution, not just the user query. The authors assume that an IR system offers a range of query expansion algorithms and retrieval algorithms. Retrieval techniques offered by the system could include 'boolean retrieval, Probabilistic retrieval, cluster-based retrieval and vector-space retrieval'. The relationships between these retrieval algorithms and query expansion algorithms are described in meta-data known to the QoS Adaptation component. This meta-data describes the way in which algorithms can be combined. When a user provides feedback on a set of results through a graphical interface, the QoS Adaptation component selects the query expansion and retrieval algorithm for the next iteration cycle. Selection of the appropriate algorithms is dependent upon the difference between the actual retrieval success rate and the agreed retrieval success rate. When actual retrieval rate is considerably below the agreed parameter, the retrieval algorithm is changed. When the actual retrieval rate approaches the agreed parameter, only the query expansion algorithm is changed.

> 'This strategy is reasonable for several reasons. Changing the retrieval algorithm implies a new relevance indicator algorithm or new vector bases for the query vector or both. Changing the query expansion algorithm means to stay with the current vector bases of the query and the documents, but to have a new query vector'. [254]

3.9 Conclusion

In this chapter the practice of relevance feedback in information retrieval systems has been examined, with particular attention to the vector space and probabilistic methodologies. The survey reveals certain aspects of relevance feedback theory that are widely endorsed by large sections of the relevant field. For instance, there seems little disagreement within the information retrieval community that the user's initial query is often a poor reflection of that particular user's information need. The overwhelming consensus is that relevance feedback represents the best method currently available for improving the quality of this query, and therefore the calibre of results a particular IR system is capable of producing.

However, beyond this point consensus breaks down. While the majority of systems improve the query using some re-weighting scheme, there is little commonality in the methods applied, and considerable divergence in relation to the relative importance of original query terms, 'relevant' expansion terms and 'non-relevant' expansion terms. Similarly, while many information retrieval systems rely on query term expansion to improve retrieval effectiveness, there is no conformity in either the number of added search terms, the quantity and characteristics of documents from which those terms are extracted, or the sorting/selection techniques set in place to ensure quality.

Given the above, there remains something of the black art to the whole process of relevance feedback which defies rapid and comprehensive analysis. It is submitted that the study of re-weighting/expansion phenomenon, whereby mathematical relationships between the various aspects of relevance feedback and system performance are proposed and then tested, is to be encouraged as a counter-measure. Also encouraging is the overdue development of IR systems which deploy a broad, multi-model range of retrieval/expansion algorithms, intended to maximise the likelihood

of satisfying any given information need.

Chapter 4

Co-active Search Techniques

4.1 Introduction

In CHAPTER 2 the problem of semantic ambiguity in relation to automatic hypertext generation systems was discussed. This was followed, in CHAPTER 3, by an examination of the relevance feedback schemes used in information retrieval to improve a set of search results. Fragments of both these surveys are now fused into a relevance feedback scheme of much broader scope, known as the *co-active* search approach. This approach was developed specifically to deal with the notorious problem of lexical ambiguity in query based search. The theoretical underpinnings of this methodology are introduced, and key principles and assumptions outlined. Examples in this chapter centre upon the search and retrieval of hypertext documents. However, as CHAPTER 6 demonstrates, co-active search can be applied to other search environments with little or no modification.

4.2 Lexical Ambiguity in Query-Based Search

As previously discussed in §2.5, lexical ambiguity is an inherent and unavoidable feature of human language, a naturally occurring characteristic of both free prose and more constrained compositions [172, 173]. This is to say that the meaning or meanings of a word can alter dramatically depending upon the context in which that word is used, so that one single word can routinely map to multiple senses. Consider,

for example, the English word *net*. Used in relation to fishermen or fishing vessels, this word conveys a set of well understood concepts, a parcel of associations and mental imagery that represents one sense of that word. However, when used within the context of computer science this word has vastly different connotations, indicating a network of interconnected computers.

Humans typically resolve word sense ambiguity[1] quite unconsciously, whether the ambiguity is part of a written dialog or some verbal communication. This resolution draws upon the immediate context of the ambiguity (what is being discussed at that moment?), combined with a ratification of which sense is the most appropriate at that given time. This process operates automatically and with such a low error rate that it is only as an aspect of modern humour that we even fully acknowledge its presence.

However, lexical ambiguity raises *considerable* problems for any retrieval system implementing query-based search. When a query contains ambiguous terms, what *exactly* is the user looking for? Initiatives in the fields of Information Retrieval, Natural Language Processing and Artificial Intelligence, aimed at this very problem, have yet to converge upon an exhaustive and practical solution. The overwhelming majority of these strategies depend, in some way or another, upon an external knowledge source to resolve word sense ambiguity. This external knowledge source has taken many forms, including expert systems [140, 138], dictionaries [122, 81], knowledge bases [140, 138, 147], sense-tagged text[40, 320] and bi-lingual parallel corpora[315]. The shared weakness of all these approaches is the time and resources necessary to develop these external knowledge sources[2]. In addition, there is the onerous administrative burden of keeping information up to date as a given domain evolves [105] (see §2.5).

4.2.1 Using the Users

Co-active search solves the problem of lexical ambiguity in query based search by utilising the power of user consensus. While co-active search is a wholly novel approach,

[1] This includes homonymy (e.g. *When I visit Brighton, I would like to see the sea*) and polysemy (e.g. *Fortunately, Bill had sufficient funds to settle the bill*).

[2] This is the 'knowledge acquisition bottleneck' discussed in [187]

inspiration is drawn from other successful systems and procedures that harness user opinion to reach operational conclusions. For example, consider the PageRank algorithm, cornerstone of the Google web search engine. PageRank has demonstrated conclusively that user opinions - in this case expressed as countable hypertext links recommending another web site - can be leveraged to gauge the likelihood that a website will satisfy a user's information need [37]. This multi-author rating or peer consensus works well because it is both reflexive in operation and distributed in its application. Anyone can increase the standing of a web site they think is helpful or relevant in any particular field with a simple anchor tag.

A further example of the power of opinion can be found in the realm of eCommerce, where the purchase of one item from an online store often elicits recommendations of related products. In this case the opinions of earlier users, as expressed in previous purchase activity culled from logs, are supplied to later users as spot or impulse products prior to checkout. If enough customers think product x is the natural bed fellow of product y then the process described above will display that information to a later user. Again, as with Google, the strength of the approach lies in enfranchising the users of a system and harnessing them in the process of selection. The popularity of this strategy with online retailers is a ringing endorsement of its success.

The co-active solution to the word sense problem is instructed by the logic described above. Rather than attempt to catalogue the unending senses a given word can exhibit (with domain modifiers, rules, ontologies etc.), word senses are determined by recording and analysing the navigational behaviour of the users of a search engine. This navigational behaviour is used to auto-categorise sets of search results, enabling the user to quickly select results matching the appropriate word sense. Co-active search records viable word senses in a consensual fashion, a process which can respond to rapid changes in the possible meanings a word may carry.

4.3 Co-active Search Techniques

4.3.1 Key Assumptions

The theory of co-active search rests upon two central assumptions, both of which concern user behaviour. Those assumptions are as follows:

1. It is reasonably assumed that whenever a user submits a query term which may carry a number of different meaning, they will tend to select resources that match just one specific meaning. For example, where a user enters the query term *ball* with the intention of locating resources about formal social events, it is unlikely that they will later select resources corresponding to some other sense of that word. We call this the assumption of *semantic consistency*.

2. Whilst the possibility of the one search - one selection transaction is not discounted altogether[3], it is assumed that the average user, searching for a given resource, will make two or more selections from a set of results during the course of a search. We call this the assumption of *search persistence*.

As these two key assumptions provide the very foundations upon which the theory of co-active search rests, we now examine them in greater detail.

[3]There are two possibilities when such a transaction occurs:

(a) The user is searching for a singular and unmistakable item. The search results are clear and unambiguous. The user selects the desired resource and the search concludes. We call this type of transaction a *naive* success, and experience tells us that its occurrence is relatively rare.

(b) The user has searched for the item before, and is aware that a query search of x will list item y in the results. The user selects the item (as they have in the past), perhaps guided to the surrogate by the navigational hints supplied by the browser (i.e. visited links often presented as a different colour from active but unvisited links). We call this type of transaction an *informed* success. Anecdotal evidence suggests navigation-by-query, as described above, occasionally supplants the book-marking strategy used to remember the location of previously viewed information.

4.3.2 Justifying the Assumptions: Semantic Consistency

The process of justifying these underlying assumptions is hampered by the sparsity of literature dedicated to the discussion and analysis of very large query logs. This rather surprising lacuna was discussed by Jansen et al. in 2000, who bemoaned the lack of complimentary data available at the time of their study:

> While there are many papers that discuss many aspects of web searching, most of these are descriptive, prescriptive, or commentary. Other than the two mentioned previously, we could not find any studies of web searching similar to this one, containing data on searches, thus we have nothing to compare our studies [152].

However, it seems fair to say that this situation has been improving steadily over the last 10 years, thanks mainly to the Excite Research Project [293] and a similar exercise carried out by Silverstein et al. [281]. The latter study reported an interesting analysis of around 1 billion entries taken from the query log of the Alta Vista search engine. These logs represented 6 continuous weeks and approximately 285 million users sessions, where a session (in this context) is understood to be an attempt to fill one particular information need. The authors found that that users of a web search engine seldom 'modify their query':

> 'Surprisingly, for 85% of the queries only the first result screen is viewed, and 77% of the sessions only contain 1 query, i.e., the queries were not modified in these sessions'[281].

These findings were later confirmed by Jansen et al. who reported on a detailed analysis of the query log of the Excite search engine [152] . Analysing just under 51,000 queries submitted by 18,113 users, the authors concurred with the Silverstein et al., finding that 'query modification was not a strong trend'. Only 22% of all queries evaluated showed evidence of query modification, where 'a modified query is a subsequent query in succession (second, third..) by the same user with terms added to, removed from, or both added to and removed from the unique query'. In this context, see also [291], which established that 'the most common user session was ...

a unique query followed by a request to view the next page of results with no query modification'[4].

These findings should be read in careful conjunction with [290], where the authors reported on the results of an interactive survey answered by users of the Excite search engine. In addition to profiling information (e.g. occupation, education level, gender etc.) 316 users provided data describing their search session. This data included details of their *current search topic*. It was found that

> '...most respondents searched on a single topic as determined by their query terms and search topic statements. Eleven respondents reported searching on two different topics and two respondents reported searching on three topics'[290].

A further 16 users in the sample searched for multiple search topics, behaviour characterised by the authors as *general information or surfing searches*. This means that just under 91% of the sample searched for just one topic during the reported search session (see also [292]).

The two findings outlined above, namely, that

1. Users of a search engine rarely modify their initial query *and*

2. Users of a search engine rarely search for more than one topic.

suggests that the typical user of a web search engine is relatively dogged in their pursuit of information. In this light, the case for an assumption of semantic consistency - which merely recognises this singleness of purpose - seems strong.

4.3.3 Justifying the Assumptions: Search Persistence

If we accept that a typical user is inclined to provide the least possible effort in exchange for satisfaction of their information need, can we then justify the assumption that they will tend to make more than one selection from the list of results provided

[4]A follow-on study of around 1 million Excite queries confirmed that there was a high degree of locality in the queries submitted to the Web search engine - 'one out of three of the queries submitted has been previously submitted by the *same* or another user' (my emphasis)[202].

by the search engine? Hopefully, Yes. To begin, the truth in this assumption was certainly hinted at in [22], where an analysis of 500, 000 clickthrough records taken from the Lycos search engine revealed 243, 595 unique queries but 361, 906 unique URLs selected by the authors of those queries (i.e. a query:URL ratio of 1:1.49).

A further clue can be found in [124] where the authors examined the 'browsing patterns of users during search tasks' by evaluating web proxy access logs. A study of 13, 657 search sessions carried by employees of the Lucent company revealed that 20% had involved the selection of 3 or 4 URLS[5].

Finally, more recent support for this contention is provided by Wen et al., who described their work with the Encarta online encyclopedia:

> 'About 1 million queries are made each week, and about half of query sessions have document clicks. Out of these sessions, about 90% have 1-2 document clicks' [313].

Provided these observations about the quantity of URLs requested in relation to queries submitted hold true, the assumption of search persistence is tacitly validated.

4.3.4 Session records

The co-active approach to relevance feedback (and specifically the problem of ambiguous queries) centres upon analysing the navigational choices of users of a search engine. Therefore, the first stage of the process is to gather data from each user search session. For each search initiated by the user a unique session id (ID) is recorded, in addition to the query term (t) that was submitted, plus the address (a) of any hypertext document that particular user selected. Each search is therefore represented by the following ternary relation:

$$(\text{ID},\ \text{t},\ \text{a})$$

[5]This percentage relates to 'completed' search sessions - 'We refer to a search session as completed if it contains at least one URL. In the remaining cases, we assume the user examined the search engine results and decided that none of the Web pages were relevant' [124]. 44% of the 13, 657 search sessions were completed.

Where the session extends to several searches the activity is recorded as a set of triples, connected via the session ID attribute:

$$\{(\mathrm{ID_x},\ \mathrm{t},\ \mathrm{a_1}), (\mathrm{ID_x},\ \mathrm{t},\ \mathrm{a_2}), (\mathrm{ID_x},\ \mathrm{t},\ \mathrm{a_3})\}$$

These session triples are stored in a database in preparation for batch processing.

4.3.5 Feature Space

All associative data is recorded in a Euclidean feature space. Each search term encountered by the system is represented as a *dimension*, and document matches to this search term are recorded as attributes upon it. Hypertext documents may exist on many dimensions (and hence relate to many search terms), but importantly possess a position upon each dimension between 0 and 1[6]. This spatial coordinate is vital as it will allow us to cluster similar documents, as will be discussed in §4.4. The set of all dimensions currently stored represents the system's feature space as a whole.

When the session records indicate that a user selected hypertext document x when searching for term t, x is assigned an initial value in the feature space along dimension D_t, denoted as x_t. The value assigned at this point is entirely arbitrary and a random value is currently used.

Where the session data reveals an *association*, so that hypertext documents a and b - which are already present in the feature space - are consecutively selected by a user searching for term t, we move attributes a_t and b_t closer together along dimension D_t. Having calculated the midpoint m of those two attributes using $m = (a_t + b_t)/2$, the actual distance each attribute moves towards m is now determined by two factors: the dimension's convergence choke and the dwell time weighting factors for each document. A visual representation of the association of a_t and b_t is shown in FIGURE 4.1.

[6]One could view each image in the collection as an implicit feature vector of length u, where u is the number of unique search terms that have been submitted to the system, but this information is not held explicitly due to issues of scalability.

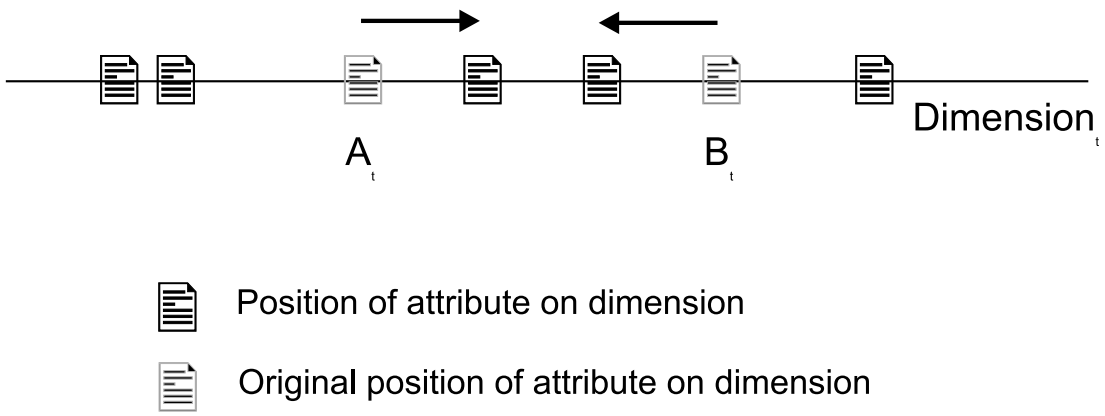

FIGURE 4.1: Effect of association between A_t and B_t

4.3.6 Dwell Time Weighting

An exception to the normal associative rules discussed above is made where a review of the session records indicate that the user spent relatively little time viewing one of the members of an association. In such a case, a dwell-time weighting scheme reduces the strength of the association or removes it completely. The dwell-time weighting component was introduced to reduce the impact of the mistaken traversal, whereby a user navigates to a document but quickly realises it does not satisfy their information need. This is an obvious problem where the short description representing the target document is insufficiently clear, and proved troublesome in early tests. In later experiments a mean dwell time was established for all searching operations which was used to weight any association. The dwell time weighting factor (DTW) was calculated using

$$\mathrm{DTW} = \frac{\mathrm{d}}{\mathrm{M}}$$

where

d is the dwell time for that search item

M is the mean dwell time for search items of this type

and the maximal value for DTW was capped at 1. A further rule for the application of DTW was established which operates as follows. Where an association is

made between n attributes, only one of which has a DTW exceeding some minimum threshold value[7], the associative operation (i.e. any movement as a result of that association) is cancelled completely.

As a worked example, assume that a user is searching a collection of documents and the average dwell time we have recorded is 19 seconds. This time represents the duration from the moment the document surrogate is selected until that moment the user returns to the list of search results. It therefore includes both the time spent weighing the document against the information need and any arrangements the user makes to capture that document or record its location (i.e saving operations, book-marking, recording the address manually etc.).

The user is searching the collection with query term t and has looked at four documents, with physical addresses of a, b, c and d respectively. All of the documents are already represented in the feature space and have non-zero values along dimension t. The dwell times for each image and the appropriate dwell time weights to apply are recorded below in TABLE 4.1.

TABLE 4.1: Dwell Time Weights

Attribute	Dwell Time (secs.)	DTW
a_t	25	1
b_t	21	1
c_t	17	0.89
d_t	2	0.11

4.3.7 The Convergence Choke

The convergence choke is a dampening factor used to slow down the movement of attributes or accelerate their attraction to each other within the feature space. Some dimensions refer to words which are searched for frequently, whilst other dimensions refer to words less commonly used as part of some query-based search. The purpose of the convergence choke is to slow down movement along popular dimensions, whilst

[7]We favoured a minimum of 0.1

simultaneously encouraging movement on less subscribed dimensions. Each dimension d in the feature space has a convergence choke value (CCV), which is calculated using

$$\mathrm{CCV} \;=\; 1 - \frac{\mathrm{n}}{\mathrm{N}}$$

where

n expresses the number of recorded associations on dimension d

N expresses the number of recorded associations on every dimension.

The convergence choke value subsequently becomes a part of the calculation to determine the distance moved by two or more associated attributes. Assuming an associated attribute a_t and a calculated midpoint of m, we obtain the modified location of a_t via

$$\mathrm{a_t} = \mathrm{CCV}(\mathrm{m} - \mathrm{a_t}) \times \mathrm{DTW}$$

4.3.8 Attribute Signatures

The relative importance of an attribute along a given dimension can be judged according to the range of movements it describes over time. Assuming an even initial allocation of attributes along the dimension, a *dominant* attribute will tend to move in both directions (i.e. positive and negative), as other attributes are drawn to it by the convergence algorithm. In contrast, a *subordinate* attribute will tend to move in just one direction, the direction determined by the position of its nearest applicable grouping. A *radical* attribute will record a relatively large movement in either a positive or negative direction, typically indicating the split or separation of a large group. A *solitary* attribute is a attribute that has not formed part of any recorded associations to date, and as such does not move along the dimension at all.

By mapping attribute movement against time and joining those points with a continuous line it is possible to identify a visual signature for each attribute type, as illustrated in FIGURE 4.2.

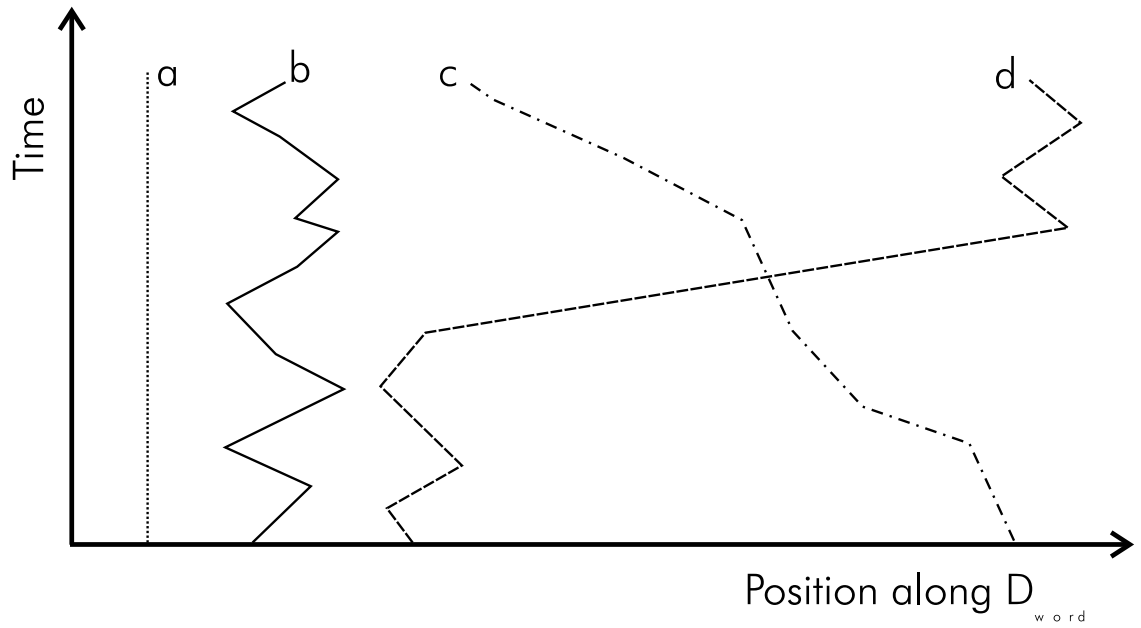

FIGURE 4.2: Attribute types: (a) Solitary (b) Dominant (c) Subordinate (d) Radical

4.4 Cluster Analysis

Given sufficient time and sufficient associations certain distributions of attributes will emerge along any given dimension. As attributes move and are attracted towards each other by the convergence algorithm discussed above, the feature space will witness the formation of semantic clusters. These clusters will represent groups of documents related to a query term used in a similar semantic sense (see §5.6.7). The next step of the procedure uses well documented cluster analysis (CA) techniques to extract these groupings in an unsupervised fashion [136, 75, 153, 79] (see also [146]).

The Euclidean distance between implicit feature vectors in the feature space provides a simple document similarity measure - the greater the distance between two feature vectors, the more dissimilar the documents. The Euclidean distance of a feature vector $a = x_1, x_2, x_3..x_n$ and feature vector $b = y_1, y_2, y_3..y_n$ is calculated as:

$$\text{distance (a, b)} = \sqrt{(\mathrm{x_1} - \mathrm{y_1})^2 + (\mathrm{x_2} - \mathrm{y_2})^2 + ...(\mathrm{x_n} - \mathrm{y_n})^2}$$

More importantly - for this particular application - the proximity of an attribute

of a feature vector along a given dimension d to the corresponding attribute of another feature vector provides a *word sense similarity measure*.

The aim of the clustering operation now performed is to group the documents into approximate categories, one category for each significant word sense. Note that at this point the clustering process is only ever concerned with one dimension of the feature space at a time, the particular dimension fixed by the query term submitted by the user. In this respect, the clustering approach used is certainly not polythetic (i.e. does not simultaneously use all attributes to calculate distances between feature vectors) although neither is it monothetic [10].

4.4.1 Clustering Algorithm

The number of clusters present along a given dimension is approximated using an agglomerative clustering technique known as the *complete linkage* algorithm [166]. This algorithm is procedurally simple and perfectly adequate for initial testing[8]. Having calculated and ordered all inter-cluster distances - with single attributes initially treated as clusters - the nearest pair is iteratively grouped to form a larger cluster. This process is continued, with the distance between two clusters measured as the *maximum* distance between any attributes from those two clusters, until a complete graph (or *dendrogram*) is formed.

The number of viable sense groupings along a dimension is now determined by examining the distance between clusters as they were connected at each stage of the algorithm. The first connection to exceed the *mean connection distance* is determined. Only clusters existing before this point are presented to the user. In this way only statistically significant groupings are formed. This automated process for determining the number of clusters means there is no need to manually set the number of senses a word should have, a notoriously difficult process [164]. This whole process is illustrated in FIGURE 4.3.

[8] [175, 10] have indicated that this approach is less appropriate for very large collections. Therefore, further experiments will utilise algorithms developed in the data mining field, where extremely large data sets are very much the norm [328, 217].

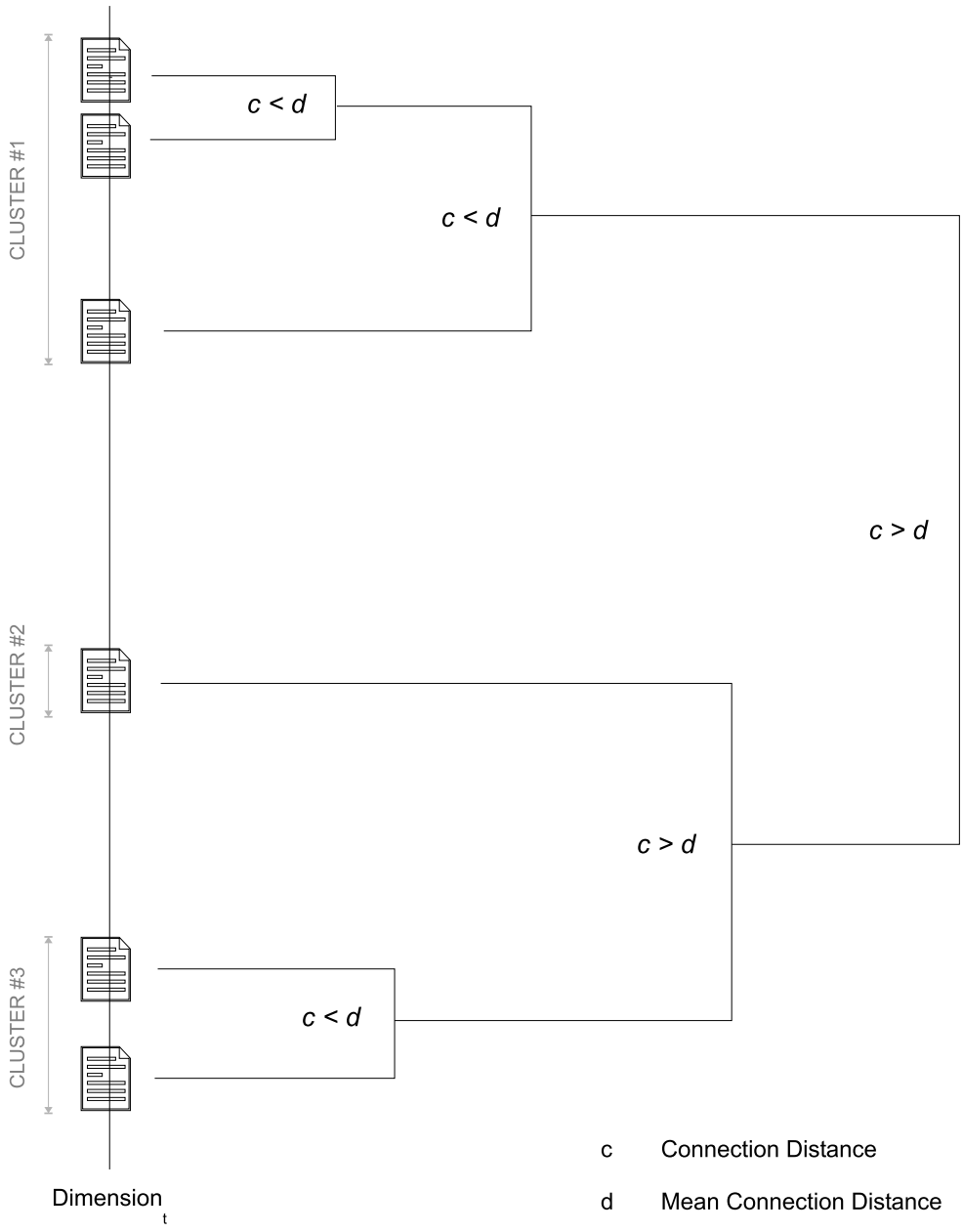

FIGURE 4.3: The SENSAI sub-mean clustering algorithm

4.4.2 Using the Sense Groupings

Having determined the cluster solution for a particular dimension, this information can now be fed back into the search process. As demonstrated in CHAPTER 5, the sense clusters derived using the techniques above are used as a filter for a set of results supplied by a search engine. Each result is automatically categorised according to the most appropriate sense category in the feature space, and the user is presented with sets of results rather than a composite list.

4.5 Related Work

The vital aspect of this approach is that it does not rely on a manually prepared source of information to discriminate between word senses. As such it is automatic and unsupervised in its operation [273] and democratic rather than autocratic. This means it contrasts favourably with those systems discussed in §4.2 which draw upon some form of knowledge source during operation.

Various authors, particularly in the field of Natural Language Processing, have discussed the problems of relying on external knowledge sources, both in relation to acquisition and maintenance, and ameliorative strategies have been developed. For example, Pedersen & Bruce introduced a corpus-based approach to word sense disambiguation that eschewed external information. Their technique relied upon only that information which could be extracted automatically from the document collection under examination - a 'knowledge lean' approach [227, 226]. Their technique utilised the Naive Bayes model [182] and the Expectation Maximisation (EM) algorithm, in both its classic [85] and stochastic formulation [117], to discriminate between word senses in an unsupervised fashion.

Similar approaches capable of inducing senses from raw text without supervised training or external knowledge sources include the context group discrimination methodology of Schütze, who interpreted senses '..as groups (or clusters) of similar contexts of the ambiguous word' [272] (see also [326]); SenseClusters, an open source, unsupervised word sense resolution system which takes '..a raw unstructured corpus, and

clusters instances of a given target word based on their mutual contextual similarities' [232]; and Fukumoto & Suzuki, who developed a word sense disambiguation algorithm based on mutual information-based (Mu) term weight learning [109];

All of the strategies above are predicated, to a greater or lesser degree, upon the Strong Contextual Hypothesis advanced by Miller and Charles [209]. This states that '..two words are semantically similar to the extent that their contextual representations are similar'. By extension, this hypothesis suggests that two senses of a given word are similar to the extent their contextual representations are semantically similar [272].

Co-active search is significantly different from the strategies above in that sense discrimination is not the product of some analysis performed on the contextual roots of the target ambiguous word, but rather an extrinsic factor. The technique capitalises on the transparent aggregation of navigational choices made by users. In this respect the co-active approach does share procedural similarities with various information retrieval and filtering technologies that have utilised what Fitzpatrick et al. termed the 'information consuming habits' of users to improve retrieval effectiveness [98] (see also [124, 212, 154, 314]).

4.6 Conclusion

In this chapter the assumptions and principles that express the co-active approach to search have been introduced. In the next chapter a concrete implementation of these theories - a system known as SENSAI - is discussed in detail. A full user trial testing the retrieval of hypertext documents is outlined, and experimental results analysed.

Chapter 5

Co-active Search: Information Retrieval

5.1 Introduction

In the previous chapter the theoretical underpinnings of co-active search were outlined. In this chapter a concrete implementation of its application to hypertext search is discussed. The SENSAI system is introduced and its various components illustrated. A regime testing the key assumption that retrieval effectiveness is increased where sense categorisation is available is outlined, and a set of experimental results analysed.

5.2 The SENSAI System

The SENSAI system is *not* an information retrieval engine. Rather, it is a set of components and technologies that are positioned *between* a user and a search engine. These components provide a sense categorisation *service* to the user, seamlessly complementing the functionality of the underlying search system

SENSAI can be 'bolted on' to any pre-existing IR tool. Although the following chapter relies heavily on the Google application programming interface (API), there is no particular reason why some other stand-alone information retrieval system (be it probabilistic, boolean, latent semantic indexing, vector space or other) could not be used in its place. Provided the IR system in question can provide initially sensible results that promote an acceptable spread of associations, the SENSAI system can

FIGURE 5.1: The SENSAI Interface

be applied.

5.2.1 The Interface

Users of the SENSAI system conduct their searching via a web form generated using the PHP scripting language. Query terms are entered into a text box by the user and the form is submitted. The query terms are then parsed by a PHP script and encapsulated within a SOAP message [32]. This message is relayed to the Google (API) [119] which replies in the same format, listing the relevant documents for that query.

Upon receipt of the reply the PHP script presents the results to the user as a list of document surrogates, modified so that each hyperlink to a Google-suggested

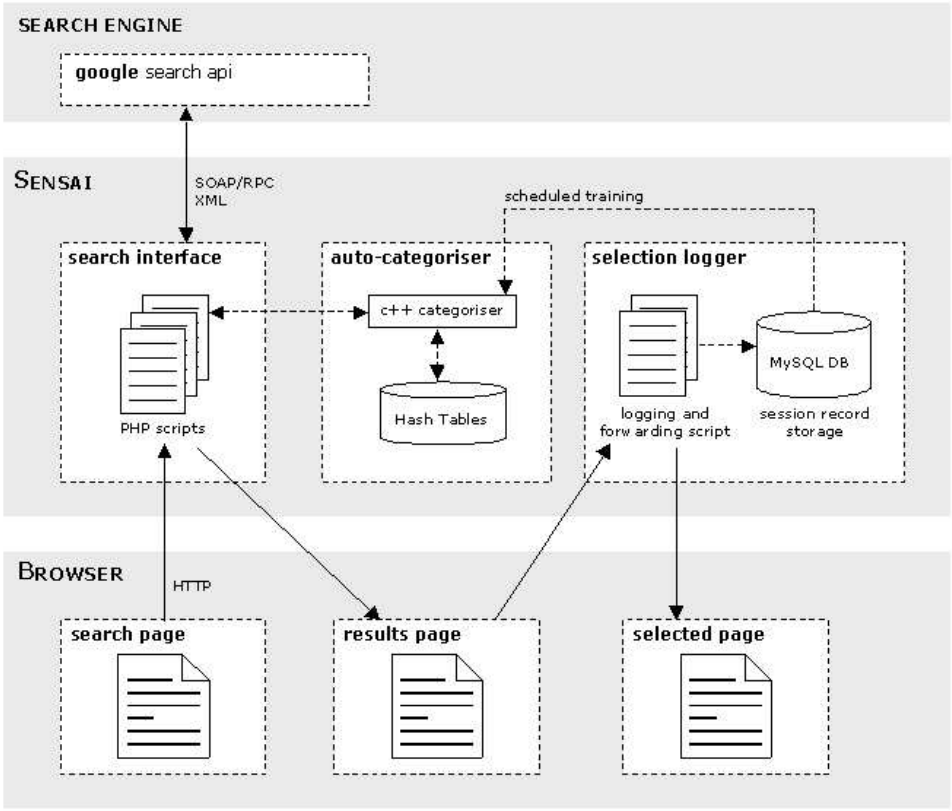

FIGURE 5.2: The SENSAI system

document will pass through a logging proxy upon its selection. As users follow selected documents from the suggested list the system transparently records their navigational choices, whilst simultaneously serving them the selected web pages. Throughout this operation URLs are masked so that the user's search process remains uninterrupted. All information relating to the user's search session is anonymously recorded in a mySQL database for later processing. Users continue their searching behaviour until their information need is satisfied or their commitment to the search wanes [126].

Note that at this time the system is still in *passive mode*. This means it is merely acting as a conduit for a search engine and is entirely ignorant of word senses (as represented in figure 5.2 through solid arrows). The system does not begin the next stage of its operational life until the captured navigational behaviours of the user clientele are analysed..

FIGURE 5.3: Categorised Search Results

5.2.2 Processing the Session Records

During periods of low demand for the system, the navigational data collected by the logging proxy is analysed and reduced to a parcel of associations modelled in SENSAI's feature space. Attributes representing documents that use any given word in a similar sense are then encouraged to converge by the algorithm described in §4.3.5.

5.2.3 Categorising the Results

When the population of the feature space reaches some agreed level, the system is ready to serve results to the user in *active* mode. Having received a live query, SENSAI derives the sub-mean cluster solution for the appropriate dimension and

feeds this knowledge back into the search results. The search coordinator script passes the list of Google-suggested documents over an internal socket connection to the c++ categorisation module. This module examines the list of documents and then categorises them according to the newly-calculated cluster solution[1]. These categorised results are then passed back to the PHP search coordinator and relayed to the user.

The net result of these operations is as follows - for a query having one search term t_1 and a sub-mean solution producing k clusters, the user is provided with $k+1$ categories of possible document matches[2]. Each category contains a grouped list of document surrogates, and the user is free to move between document categories (and therefore between senses of the query) by clicking on a representative thumbnail. These sense categories represent a human-intuitive overview or survey of the distinguishable semantics of the query, like grouped with like [207].

5.2.4 Abstracting the Category

To guide user navigation and selection each of the sense categories served to the user is abstracted. This *semantic signpost*, which appears next to the category thumbnail, is created automatically using the following procedure:

1. Each of the documents belonging to a single cluster is parsed and all HTML [1] mark-up is removed.

2. All of the parsed documents are then concatenated into one large text file.

3. A stop-list is applied to this text file, removing common words such as propositions, conjunctions and pronouns. The stop list used for this particular application is Fox's formulation for general text [103, 102], shown in APPENDIX A.

4. The text file is then alphabetically sorted, and the frequency of each word is scored.

[1] Where a Google-suggested document does not appear in our feature space it is placed in a special category for *unclassified* documents, and a new (implicit) feature vector is created to represent it.

[2] The extra category contains unsorted documents.

5. The ten words with the highest frequency are then used as the abstract to describe that particular category[3].

5.3 Additional System Functionality

The following section describes additional features that were added to the SENSAI system during implementation. They do *not* form part of the core theory of co-active search but *do* illustrate its flexibility and ease of extension.

5.3.1 Search Modes

SENSAI offers two search modes. The *iterative* search mode will be the mode most familiar to typical users. In this mode a user repeatedly searches the collection, modifying or replacing the query at random, selecting and reviewing candidate documents, until eventually their search is complete or abandoned. The only difference between the system in this mode and a traditional hypertext search engine is:

1. The transparent gathering of session data, to further inform the feature space.

2. The automatic and flexible categorisation of word sense.

The *regressive clustering* mode, on the other hand, enables the user to further subdivide a cluster of sense-similar documents in order to 'focus in' on the correct word sense [9]. By selecting the abstract of a category instead of a document surrogate from the results list, the user forces the complete link algorithm one step further back in its tracks, with the spatial focus on the cluster that abstract represents. This results in a sub-dividing of the cluster into its nested constituents, yielding a new set of surrogates and cluster abstracts. The user can continue this process until the progressively finer granularity of the sense clustering yields appropriate results. This process is illustrated in FIGURE 5.4.

[3]A similar effect could possibly be obtained by abstracting the centroid of the sense cluster [92, 261]. The centroid is a dummy feature vector obtained by averaging the selected feature vectors in the cluster [266].

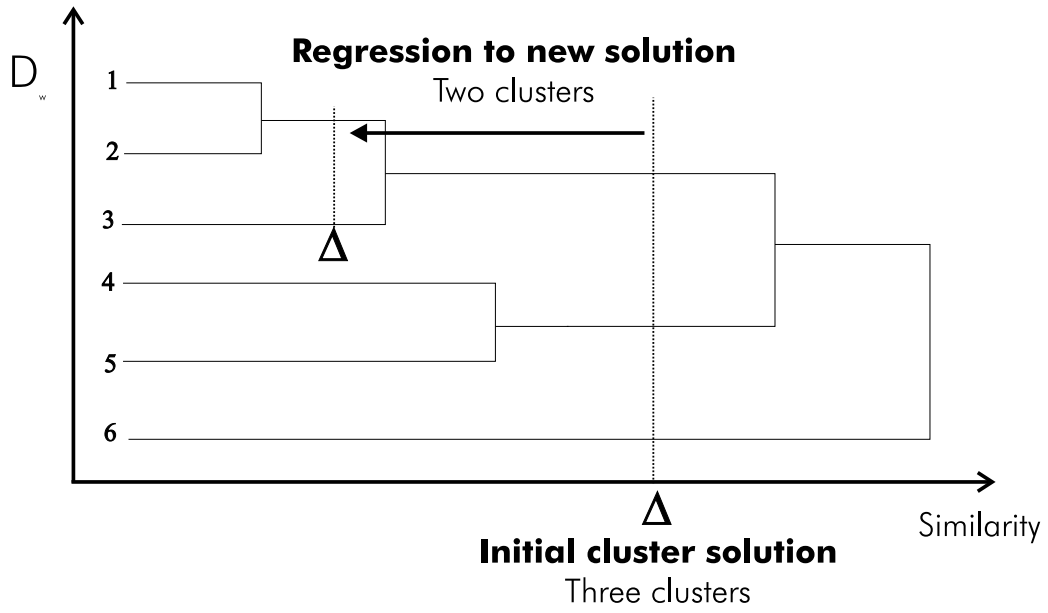

FIGURE 5.4: Dendrogram regression enabling focussed searching

5.3.2 Negative Associations

Where an attribute is incorrectly clustered so that a document appears in the wrong category, SENSAI offers the user the opportunity to rectify that mistake by providing negative feedback. When this component is enabled, each document surrogate has a small icon adjacent to it displaying a question mark symbol. By clicking on that icon the user is stating that the identified document is inappropriate in the context of the current category. This click-through behaviour is captured by the search logger, and the attribute involved is moved away from the nearest cluster, the distance moved determined by a fixed system value in tandem with the convergence choke. The user is then served an amended list of suggestions with the nominated document absent.

Negative feedback allows the system to account for those attributes that have been allocated a position along a dimension which coincidentally is the focus of a unrelated high density cluster. If a number of users follow the procedure outlined above then even solitary attributes (which normally remain fixed at one point along a dimension) can be 'bumped' out of a category.

5.3.3 Caching

The module responsible for automatically categorising any set of results provided by the Google API is updated periodically with the latest set of session records. Between updates, when the feature space is in a constant state, there is an opportunity to cache cluster solutions for common words, thereby improving system turnaround when a user enters a query.

Caching is a background process which is interrupted by the submission of a live query. The caching algorithm attempts to calculate a cluster solution for every dimension present in the feature space, with precedence given to heavily populated dimensions. When a new query is submitted to the system it is checked against the list of pre-compiled cluster solutions. If a match is made, that solution is served to the user. If no match is present in the cache list, the solution is calculated immediately, passed on to the client, and stored for later use.

The cluster solutions obtained using dendrogram regression can also be cached by the system, but have less general application. This is because they represent a quite specific information need unlikely to be repeated elsewhere.

5.4 Testing the System

In this section the hypothesis that efficient word sense disambiguation will ultimately improve retrieval effectiveness is finally tested [172, 173][4]. The metric used to test this assertion is *total search duration* (i.e. the time taken by the test user to complete a number of search tasks). Several other performance indicators could have been applied (see [245, 237]), but total search duration was judged to be the most appropriate.

5.4.1 Problems of Scale

SENSAI is a meaning resolution system that relies upon its users, and whilst this is an inherently good thing, for all the reasons discussed in CHAPTER 4, it did present

[4]This hypothesis is in direct opposition to Sanderson's findings on the negligible effect of word sense disambiguation on retrieval effectiveness as a whole [269]. However, it is worth considering later observations pointing towards the authors selection of pseudo-word [273].

certain technical challenges. Chief amongst these was the immediate need for a large user base to test the system in its early stages. It was frequently observed that the quality of automatic categorisation was directly related to the number of active users of the system. This is perfectly logical. The greater the number of users, the more associations are made and recorded by SENSAI, and the stronger consensus becomes. As more associations are fed into the categorisation module with respect to a given query, the accuracy of the modelling of the semantics of that query within the feature space increases. Therefore, practicalities of the system demanded immediate large scale testing.

However, several factors were in favour of initially small user tests. Firstly, there was the ingrained logic of early experiments (i.e. start small and work up). Secondly, the effort involved in procuring a large set of user session records in the form we required would be non trivial, in all liklihood raising various issues relating to data privacy and proprietary information. Thirdly, as the number of users and documents increases, so too would the complexity of analysing the fundamental workings of the feature space.

The solution to this formidable issue of how to test a large scale application using small scale resources is detailed below. In brief, it consists of controlling the search vocabulary available to the user and simulating both the information need and the search results.

5.4.2 Query Allocation

As a first step, twenty test users were recruited and assigned to one of two groups, labelled *group a* and *group b* respectively. Each test subject was given a list of 10 ambiguous single word queries (known as *data set 1*), and instructed to search the SENSAI system using those queries. The queries that made up *data set 1* are shown in TABLE 5.1

TABLE 5.1: Data Set 1

apple	flush	skate
tube	bat	ball
fence	bowling	spirits
	lock	

5.4.3 Search Agendas

Each query word in *data set 1* was accompanied by a *search agenda*. The search agenda component was introduced into the testing process for two compelling reasons. Firstly, by providing the user with a hypothetical information need the test more accurately models the typical search task. Secondly, earlier experiments where no information need was assigned to the users proved disappointing, with selections heavily concentrated on the more obvious sense for each ambiguous word. While this provided limited data for clustering operations, less obvious or minority senses were occluded. This meant the ensuing categorisation task was complicated by a higher degree of noise than anticipated. Ultimately, the purpose of agenda distribution is to spread the associations across a wide range of senses in order to simulate larger scale participation.

The search agendas were constructed in the following way. Each single word query was associated with n different search agendas, where n represents the number of word senses manually established for that particular query[5]. These directives were stored in a separate table of our systems database and issued to the test users on a randomised, per-query basis. An example set of agendas, written for the query term *cavalier*, can be found below.

> *You are thinking about buying a used car (a Vauxhall Cavalier). Find three web sites that may help you in this task. ...*
>
> *Someone has just accused you of having a 'cavalier' attitude. Find three dictionary definitions of this word. ...*

[5]See APPENDIX B for a full list.

You are writing an essay on the Cavalier army that supported King Charles I. Find three web pages that might be of interest. ....

You would like to buy a Cavalier King Charles Spaniel. Find three web sites that list breeders, kennels or other likely contacts.

After submitting their one word query, users considered the search agenda allocated to them and were asked to click on the most relevant documents that SENSAI returned. This action navigated the user away from the search results page and towards the document indicated by the representative surrogate.

If this document satisfied the search agenda, users were asked to make a note of its address and return to the search results using the BACK button of their browser. Users were free to view any number of documents in this fashion, and were subjected to no time limit. However, when three or more documents had been viewed in this fashion a button appeared allowing the user to move to the next search task.

This process composed STAGE I of testing, as indicated on Figure 5.5. The time taken by each test subject to complete the ten search tasks was recorded, and an average completion time established for each test group. This control data was used at a later point to normalise a second set of results (see §5.5).

5.4.4 Building a Results List

SENSAI is a fully operational system and has access to every document suggested by Google. However, it was vital during the experiment to constrain the results list. If users had been given a very wide range of documents to consider, the likelihood of clear associative data would have been drastically lowered, impacting directly on the formation of useful clusters in the feature space. In order to concentrate user activity in specific areas, *a pseudo-results list* was created for each one word query. This artifice was implemented as a server side filter between the searcher and the Google search engine, so that only pre-approved document surrogates ever reached the user. Each simulated set of results was designed to mimic/exaggerate the constitution of the genuine article, so that each manufactured list contained:

TABLE 5.2: Data Set 2

port	tick	bug
fly	net	memory
chips	pen	paris
	cavalier	

1. *Possible Document Matches* - a number of documents that *might* satisfy the hypothetical search agenda. If a query was accompanied by n search agendas, the simulated set of results would contain n groupings of documents related to that query word, each with a different semantic.

2. *Near Noise* - a number of documents obliquely related to the search word but not strictly intended as search agenda solutions. These files represented the near misses frequently produced by a hypertext search engine, being web pages not obviously related to the query but affiliated to one of its senses in some way. For example, where the query term was *bat* and one of the accompanying search agendas was related to the sport of baseball, a near noise representative might be a web pages describing the design and construction of a baseball mitt.

3. *Noise* - a number of documents irrelevant to the query which were included to better simulate the typical search. These documents were selected at random.

To simplify the user's task, the list of simulated results never exceeded 25 document surrogates, listed 10 per page. As a precaution against prompting the user, the order in which the surrogates appeared was randomised to reduce the likelihood of rank precedence affecting document selection. Any dead links served by Google, whereby a document surrogate linked to a non-existent web page, were purposefully left unmodified as an integral part of the simulation.

5.4.5 Measuring SENSAI's Impact

Collection of control data was now concluded, and STAGE II of the experiment undertaken. A second set of 10 one word queries (plus agendas) was designed (known

as *data set 2*, see TABLE 5.2). These queries were issued to a public group of users who did not participate in STAGE I of the test. The purpose of this exercise was to populate SENSAI's feature space, thereby enabling the system to provide categorisation facilities for our core testers in STAGE III. With the convergence choke for each dimension manually set to a relatively high value of 0.9, a workable spread of attributes was obtained.

5.4.6 Active vs. Passive

In STAGE III of the trial, one of the test groups from STAGE I was required to work through *data set 2* with SENSAI still in passive mode (i.e no categorisation information available). The remaining group was also asked to work through *data set 2* but with SENSAI operating in active mode on the feature space created by the general public in STAGE II.

As the two test groups worked through the new set of search tasks, the time taken by each individual to complete the exercise was recorded and an average time established for each group. These average group times were then normalised using the control data obtained in STAGE I. A graphical representation of our testing regime is shown in FIGURE 5.5.

5.5 Results

5.5.1 Data Set One

Each test user was assigned a unique identifier when beginning the test. Users with an identifier between the range of 1–10 were assigned to *group a*. Test users with an identifier between the range of 11-20 were assigned to *group b*. TABLE 5.3 shows the completion times for all users completing *data set 1*. TABLE 5.4 presents the average completion time for each group, and the normalisation factor derived when one average group time was compared to the other.

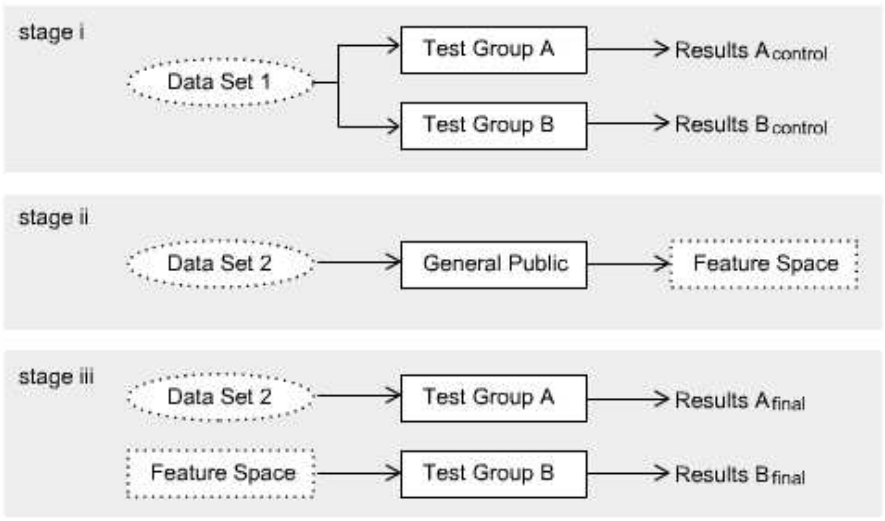

FIGURE 5.5: Testing regime for SENSAI system

5.5.2 Data Set Two

The timings for *data set 2* were recorded in a similar fashion. In this part of the exercise *group a* completed the search tasks without sense categorisation whilst *group b* worked through the search problems with categorisation information enabled. TABLE 5.5 shows the completion times for all users in seconds. TABLE 5.6 presents the average completion time for each group, in both the raw and normalised form.

5.6 Data Analysis

As can be seen from TABLE 5.6, the hypothesis that retrieval effectiveness is increased where sense categorisation is available to the user is *not proved* by the observed results. In fact, after normalisation, the average search duration recorded by Group B (*sense categorisation on*) was roughly 1% higher than the average time recorded by Group A (*sense categorisation off*). The following section analyses the various factors that may have contributed to this finding.

TABLE 5.3: Completion Times: Data Set 1

User ID	Group	Start Timestamp	End Timestamp	Completion Time (sec.)
1	A	1115211969	1115212907	938
2	A	1115212569	1115213497	928
3	A	1115222651	1115223734	1002
4	A	1115221949	1115223350	1401
5	A	1115299050	1115299811	761
6	A	1115425745	1115426275	530
7	A	1118157001	1118157746	745
8	A	1115219805	1115220452	647
9	A	1116198961	1116199401	440
10	A	1116255222	1116256453	1231
11	B	1116344495	1116346042	1547
12	B	1118068532	1118070120	1588
13	B	1118060513	1118061286	773
14	B	1118315254	1118315876	622
15	B	1118135443	1118136318	875
16	B	1115597341	1115597984	643
17	B	1118153167	1118154189	1022
18	B	1115218893	1115219609	716
19	B	1118222013	1118222938	925
20	B	1118220207	1118220752	545

5.6.1 Completion Difficulties

Of the 23 post-graduate students that were recruited for STAGE II of the trial, only 65% completed the full set of 10 search tasks (see TABLE 5.7 for details, in which a zero value in the END TIMESTAMP column indicates non-completion). While non-completion of the assigned tasks by a number of test subjects would have no particular effect on the feature space, other than a lower number of associations than originally anticipated, it raises the possibility that the test subjects recruited for STAGE II found the exercise considerably harder than the individuals participating in STAGE I.

5.6.2 High Total Search Durations

The speculation that STAGE II testers found the trial more difficult than their STAGE I counterparts is strengthened by analysis of the total search durations recorded by the

TABLE 5.4: Average Completion Times and Normalisation Factor

	Average Completion Time (sec.)	Normalisation Factor
Group A	862.3	1.073408327
Group B	925.6	1

TABLE 5.5: Completion Times: Data Set 2

User ID	Group	Start Timestamp	End Timestamp	Completion Time (sec.)
1	A	1119907277	1119907782	505
2	A	1119549300	1119549977	677
3	A	1119605159	1119605848	689
4	A	1119753243	1119754090	847
5	A	1119970184	1119971041	857
6	A	1119619183	1119619699	516
7	A	1119599208	1119599797	589
8	A	1119963737	1119964462	725
9	A	1119613837	1119614343	506
10	A	1119593156	1119594128	972
11	B	1119611410	1119612251	841
12	B	1119926620	1119927565	945
13	B	1119890265	1119890807	542
14	B	1119962974	1119963456	482
15	B	1119964502	1119965103	601
16	B	1119549266	1119550028	762
17	B	1119585662	1119586749	1087
18	B	1119567398	1119568344	946
19	B	1119611744	1119612331	587
20	B	1119906532	1119907197	665

post-graduate students. The average completion time for all students who completed the ten tasks was 23987.93 seconds (over six and half hours). The reason for this unexpectedly high average completion time can be found in the first row of TABLE 5.7. Student #1 began the test on Thu, 9 Jun 2005 14:10:59, and did not complete the exercise until Mon, 13 Jun 2005 12:46:38. In fact, it seems likely that this individual began the test, became distracted or annoyed, then completed his/her work after a restful weekend! Even if the timings recorded by Student #1 are removed entirely from the calculations, the average completion time (1377.14 seconds) is *still* unexpectedly high w.r.t the like for like exercises completed by the STAGE I testers.

TABLE 5.6: Average Completion Times w/ Normalisation Factor Applied

	Average Completion Time (sec.)	Normalisation Factor	Normalised Completion time (sec.)
Group A	688.3	1.073408327	738.8269512
Group B	745.8	1	745.8

Since the total search durations recorded by each of the students was not a relevant factor in the testing regime, these unusually high figures should have had little effect on the overall results. However, it seems likely that the difficulties the students experienced (demonstrated by their timings) fed directly into their navigational behaviour, which was of a particularly low quality.

5.6.3 Low Quality of Associations

The testers recruited for STAGE II of the SENSAI test were expected to behave in a certain way. When issued with a search agenda that addressed one specific sense of a given query word, it was anticipated that they would select hypertext documents from the results list that matched that particular semantic. However, this was not the case.

When the pseudo-results lists were constructed, each hypertext document was assigned a sense identifier. If the identifier for a document was between the range of 1-3, that document was intended as a possible solution to one of the three search agendas issued. The number 4 corresponded to near noise documents, and the number 5 indicated pure noise. Using these sense identifiers, a simple check of the SENSAI log data was carried out. Some of the users behaved consistently with regard to word sense and made predictable associations. For example, this is the output generated when SENSAI processed query term *memory* from session (28e4e5a8a974892537eca2ae3e15dfa5)[6] in verbose mode:

```
memory

association:  www.sciencedaily.com/releases/2003/11/031112073642.htm [0.22257277] [3]
```

[6]See APPENDIX C for the full transcription,

TABLE 5.7: STAGE II Timings

Student ID	Start Timestamp	End Timestamp	Completion Time (sec.)
1	1118326259	1118666798	340539
2	1118326351	1118327815	1464
3	1118326776	1118327419	643
4	1118327502	1118329058	1556
5	1118329346	0	
6	1118331211	1118331842	631
7	1118335499	1118336445	946
8	1118339330	1118340921	1591
9	1118346496	0	
10	1118355994	1118357553	1559
11	1118358743	1118359962	1219
12	1118401866	0	
13	1118404171	0	
14	1118414550	1118416317	1767
15	1118656094	1118656957	863
16	1118660302	1118661819	1517
17	1119025264	0	
18	1119018875	1119023078	4203
19	1118669929	0	
20	1118682716	0	
21	1118824721	1118825481	760
22	1119008752	1119009313	561
23	1119254324	0	

```
association:  www.medicalnewstoday.com/medicalnews.php?newsid=15950 [0.1649] [3]

association:  www.cat.cc.md.us/courses/bio141/unit3/humoral/antibodies/memory/memory.html [0.29741] [3]

association:  www.emoryhealthcare.org/press/ehc/2004/Nov/Long-term-Immune-Memory-Cells.html [0.2824] [3]
```

Notice the bracketed number at the end of each system message marked `association`, which indicates the sense group that the hypertext document concerned belongs to. In this case, the user selected four documents from the results list, all of which corresponded to sense 3 of the query term *memory* (*Find out more about the term 'memory' in relation to immunology. Find three web sites with useful information.*)

The navigational behaviour illustrated above can be contrasted with the actions recorded by another student during session (3706a7bf6d39107cf27ebb7be7b76afa):

```
net

association:  www.walthowe.com/navnet/history.html [0.63680857142857] [3]

association:  www.davesite.com/webstation/net-history.shtml [0.63680857142857] [3]

association:  www.nethistory.info/ [0.6505] [3]

association:  www.internetvalley.com/intval.html [0.6505] [3]

association:  www.isoc.org/internet/history/ [0.63680857142857] [3]

association:  www.investorwords.com/3259/net-profit.html [0.64698586666667] [2]

association:  www.investordictionary.com/definition/Net profit.aspx [0.63749314285714] [2]

association:  dict.die.net/profit/ [0.6505] [2]

association:  www.hll.com/HLL/investors/glossary.html [0.63822335238095] [2]

association:  www.wordwebonline.com/en/NETPROFIT [0.63822335238095] [2]
```

Here, the student concerned has deviated strongly from expected behaviour, selecting 10 hypertext documents associated with both sense 2 *and* sense 3 of the query term *net* (which address the history of the Internet and the concept of net profit respectively). As a result, all ten documents are parceled together into an erroneous association. Obviously, this has a direct effect upon the formation of clusters in the feature space, which in turn impacts negatively on the total search durations recorded in STAGE III - less accurate sense clusters equating to longer periods searching lists).

5.6.4 Statistical Analysis

The example given above of an erroneous association is by no means isolated. Of the 135 associations recorded by SENSAI during STAGE II trials, 28 (or 20.74%) mixed two or more of the primary senses (i.e. 1, 2 or 3). The frequency of associations involving near-noise documents was closer to initial predictions, accounting for 39 of the recorded association (or 28.89%). However, the frequency of associations involving pure noise documents was unreasonably high - 16 (or 11.85%) of the 135 associations fed into the system's feature space contained utterly irrelevant documents. In session (7872c7ed1fae4c24f22d2592a2ad7f50), one single association for the query term *pen*

contained three pure noise documents!

If an optimal association contains documents corresponding to one single sense group and no near-noise or noise documents, a post-mortem of STAGE II of the trial reveals only 61 instances. This means that over 54% of all associations made at this juncture of the experiment were sub-optimal, to a greater or lesser degree.

5.6.5 Possible Reasons

The poor performance of the testers involved in STAGE II of the user trials has to be explained, because it runs in direct opposition to the assumption of *semantic consistency* outlined in 4.3.1. In no particular order of precedence, the possible reasons for the low quality of data harvested during this portion of the trial are:

1. *Fiscal reward* - A financial incentive (in the form of entry in a cash prize draw) was offered to encourage the post-graduate students to participate in the test. It is possible that some of the testers made random or poorly considered selections in order to complete the tasks as soon as possible and guarantee their place in the lottery.

2. *Unfortunate demographics* - A very high percentage of the post-graduate students that took part in STAGE II of the test speak English as a second language. Thus it appears that some of the students may have used the web pages to help them 'understand' the search agenda, rather than the other way round. Ironically, some of the search agendas may also have been ambiguous to a non-native speaker of English.

3. *Completeness* - A proportion of the students, aware of observation, wanted to perform well. This seems to have entailed finding *all* the relevant web pages for a task, and 'making sure' other pages were not relevant.

Further testing will quickly reveal whether these results are anomalous.

5.6.6 No Dwell Time Weighting

A further problem encountered during the trial - which was not related to STAGE II testers - was the absence of dwell time weighting (see §4.3.6). This trial was designed to test the fundamental algorithms of SENSAI. As a result, much of the extended functionality of the system was deactivated before the test was initiated. One of the components that was deactivated was the module responsible for the weighting of associations based on time criteria. This suspension was deemed appropriate because the system was about to process a *simulated* search task rather than the real thing. In this rather unusual context, it was assumed that dwell time would be a less reliable metric for assessing the value of a hypertext document w.r.t. an information need.

However, this operational decision was to cause difficulties in relation to the cluster abstraction algorithm. A high frequency of the links served by Google corresponded to resources no longer in existence (i.e Error 404)[7]. Isolated instances of dead links caused no real perturbation, but problems did occur when *every single document* in a particular sense category generated the same error, leaving SENSAI no text to abstract.

A good example of this fault in action can be found by examining the abstracts generated when D_{net} was processed (see TABLE 5.8):

TABLE 5.8: Automatically Generated Abstracts - Query Term 'net'

Cluster	Abstract
1	fish fishing species sea nets water size trawl catch tuna
2	bank bn deutsche barclays uk ? business investment sell estate
3	
4	internet history web profit information were computer research arpanet world

Cluster #1 is sound, corresponding to sense 1 of the query term (*You are writing an essay on boat fishing techniques involving nets. Find three useful web sites.*). Clusters #2 and #4 are also appropriate, relating to senses 2 and 3 respectively (the

[7]The existence of dead links in a Google results list was acknowledged before the test was commissioned (§5.4.4). The decision was made to leave them *in situ* as a valid part of the search simulation.

definition of net profit, and the history of the Internet). However, the abstract for cluster #3 is completely blank. All three documents belonging to this cluster had either been moved or deleted by the time STAGE II took place. Assuming SENSAI's feature space reflected the very low dwell times spent viewing these three missing documents, it is likely that cluster #3 would have been removed from the results altogether. Dead links and obviously wrong selections were in fact the primary justification for developing the time based weighting scheme. Analysis of this fault therefore simply confirms a need for a dwell time weighting scheme which was known prior to the event.

5.6.7 Success in Adversity

Despite the low quality of data fed into its feature space, SENSAI did achieve small measures of success in relation to certain queries. For example, these were the cluster abstracts generated when D_{port} was processed:

TABLE 5.9: Automatically Generated Abstracts - Query Term 'port'

Cluster ID	Abstract
1	abp ports british associated uk southampton information june operations csr
2	wine px wines recipes ports cooking sauce color vintage chicken
3	data signal bit usb serial ports parallel computer device line

All three cluster abstracts in TABLE 5.9 map well to the search agendas (which variously addressed maritime ports, port wine and computer ports respectively). For illustration, the hypertext documents making up these clusters are shown below:

CLUSTER 1

ASSOCIATED BRITISH PORTS HOME PAGE
... is the UK?s leading ports business, providing innovative and high-quality port ... In addition to ports located all around the UK, ABP has a number of .....
www.abports.co.uk/

PORT OF SOUTHAMPTON
... Southampton is the UK?s leading vehicle-handling port, and has long been the UK?s ... www.bramblemet.co.uk - For weather information from Bramble Bank .....
www.abports.co.uk/southampton/

FERRY CROSSING CROSS CHANNEL FERRIES DOVER CALAIS MARINA

Details of the ferry port, cruise terminal and marina, plus information for passengers. ...
www.doverport.co.uk/

University of Portsmouth — home
Official site with information about the courses, departments, and applying....
www.port.ac.uk/

Ports Transport - UK Ports web sites & information - Transport UK
Ports uk web sites, local and national uk Ports information....
www.dotukdirectory.co.uk/ Transport/Ships_and_Shipping/Ports/

Medway Ports - Official Website
The Mersey Docks and Harbour Company operates the UK's second largest group of ports the Port of Liverpool, Medway Ports and the Port of Heysham ... ...
www.medwayports.com/

CLUSTER 2

port; Porto; Port DOC Definition in the Wine Dictionary at ...
... A second label vintage port, like a traditional one, is made from the better wines from various sites. GRAHAM'S Malvedos is an example of a second label ... ...
www.epicurious.com/drinking/ wine_dictionary/entry?id=7570

Port Wine Styles — Feature
... Port is a fortified wine made with a blend of many grape varietals (up to 80 ... valued wines in the world is Vintage Port with only a 2supply made...
www.cellartastings.com/en/wine-feature-port-wine.html

Introduction and main page of - The Port Wine Appreciation Pages
A comprehensive guide to Port Wine, it's production, storage, history and consumption. Including Vintage guides, Vintage Ratings, glossary, etiquette, ...
www.the-port-man.fsbusiness.co.uk/

winemaking: Requested Recipe (Port Wine)
A recipe for homemade port wine. ... Second, port is a fortified wine and this recipe uses brandy as a fortifying agent; do not use flavored brandy. ... ...
winemaking.jackkeller.net/reques13.asp

Port Wine About Port Wine. ... bullet, Draw off - to drain the juice from the tanks in which the wine is made, leaving the pomace behind. ... ...
www.goodcooking.com/portinfo.htm

ENJOYING PORT WINE - Into Wine
All about port wine including the History of Port, the Douro Valley, Port wine styles, Vintage Port, from harvesting to the lodges, Port producer profiles .....
www.intowine.com/port.html

Cooking with Wine Recipes - Home Cooking
Cooking with Wine Recipes. Wine can be used as a flavoring, as in wine jellies or in ... with Roasted Peppers, Spinach, and Goat Cheese with Port Wine Sauce ... ...
homecooking.about.com/library/archive/blwineindex.htm

CLUSTER 3

Untitled Document
... I will provide some detail in this month's tip on what computer ports enable ... software programs use a computer port to send data to another machine....
www.detto.com/learningcenter/MLCTOTMoct2003.htm

HOWSTUFFWORKS "HOW USB PORTS WORK"
... ways of connecting devices to your computer (including parallel ports, serial ports and special cards that you install inside the computer's case),...
computer.howstuffworks.com/usb.htm

COMPUTER PORT LINKS
Linking you to various locations on the Internet which have information about computer ports. ...
www.computerhope.com/network/ports.htm

COMPUTER PORTS FOR CONNECTING DEVICE PERIPHERALS
pictures and descriptions of various ports used for connecting device peripherals such as adaptive equipment and assistant technology....
www.abilityhub.com/information/ports.htm

CTIPS.COM - RS-232 SERIAL PORT
RS-232 was created for one purpose, to interface between data terminal ... The serial port takes 8, 16 or 32 parallel bits from your computer bus and .....
www.ctips.com/rs232.html

As can be seen from the document surrogates listed above, SENSAI's clustering of the results for query term *port* is highly accurate, presenting three solid groups of similar sense documents with very little inter-cluster noise. This success reflects the slightly higher quality of the associations involved. Of the 15 associations made by testers searching for the term *port*, 10 were optimal, 1 mixed the primary senses, 3 contained near noise documents, and only 1 contained noise). Other query terms with a relatively high proportion of optimal associations (for example, *chips*) also fared well. As would be expected, the query terms with the lowest frequency of optimal associations yielded poor abstracts and confused sense clusters.

5.7 Conclusions

Although the hypothesis that total search duration would be reduced where appropriate sense categorisation was available was not proved in this chapter, neither was it disproved. The trial as a whole was adversely affected by unexpected or irrational user behaviour, and is not conclusive either way.

However, the trial as a whole was certainly not a wasted exercise, suggesting many further avenues for future research and refinements to existing procedures. Additionally, even in the face of adversity, SENSAI demonstrated limited success with isolated queries, validating core co-active theories.

One very important realisation *was* made, and it is this - a simulation of a search, no matter how well constructed, is *not* equivalent to the real-world search of an information space. The systematic evaluation of the SENSAI system now demands real user behaviour, rather than some laboured simulacrum.

Chapter 6

Co-active Search: Image Retrieval

In this chapter it is demonstrated that co-active search techniques can be applied to search environments other than hypertext with little or no modification. The SENSAI approach to word-based image retrieval is introduced, and the *semantic gap* phenomenon peculiar to content-indexing systems discussed. Finally, related work in the field of image retrieval is examined for similarity.

6.1 Introduction

Two advances have rapidly changed the character of the World Wide Web: the proliferation of devices capable of effortless image capture and the widespread availability of high volume, cheap storage media. As more and more authors publish their image collections as a matter of routine, the fluid ratio of textual components to image objects in the typical web page has been subject to a considerable shift. The technical challenges of providing timely and efficient search and retrieval operations over this enormous distribution of images rivals, if not exceeds, those documented in relation to full text search.

The problem space for image retrieval is as follows: Given a query, the search system must first establish *exactly* what the user is looking for. This is no trivial task given the typically low quality of user's querying techniques and the frequent ambiguity of the queries themselves [77, 168]. Parallel to this task, the search system must also determine the semantic content of the images, an intractable exercise given

the subjective nature of image interpretation [277, 276]. The resolution of *meaning*, with all that implies, is currently an insoluble problem at both ends of the search operation. The SENSAI system slots neatly into this problem space and offers a possible solution. The application of collective intelligence enables the resolution of lexical ambiguity in query based image search whilst simultaneously addressing the determination of image content.

In the following section, the adaptations made to the SENSAI system to prepare it for use in image retrieval are discussed.

6.2 Adaptation of the System

The interface through which the user can access the SENSAI image retrieval system is identical to that discussed in CHAPTER 5. Having logged on to the SENSAI web page, the user must select the *Images* toggle from the media bar. The user then enters a query which describes the type of image he/she is searching for. The search terms entered are parsed by SENSAI's PHP scripts and submitted to the Google image search engine, which replies with a series of URLs corresponding to images that might satisfy that information need[1].

As with hypertext search, SENSAI initially operates in *passive mode*. Therefore, the Google suggested images are simply channelled to the user. Each result is represented by a thumbnail, which acts as a link anchor to the source image. The URL for each of these source images is automatically modified so that any selection made by the user passes through a logging proxy. As users select and view the images listed on the results page, their searching behaviour is recorded. This data is modelled in a feature space, with each of the implicit feature vectors present in this space representing one single *image*. Similar sense attributes belonging to these feature vectors are encouraged to move closer to each other along a dimension by the convergence algorithm operating in combination with the dimensional choke and a dwell time weighting scheme.

[1]Note that in the case of image retrieval we do not use the Google API. Instead we perform a screen-scraping operation to gather the relevant URLs.

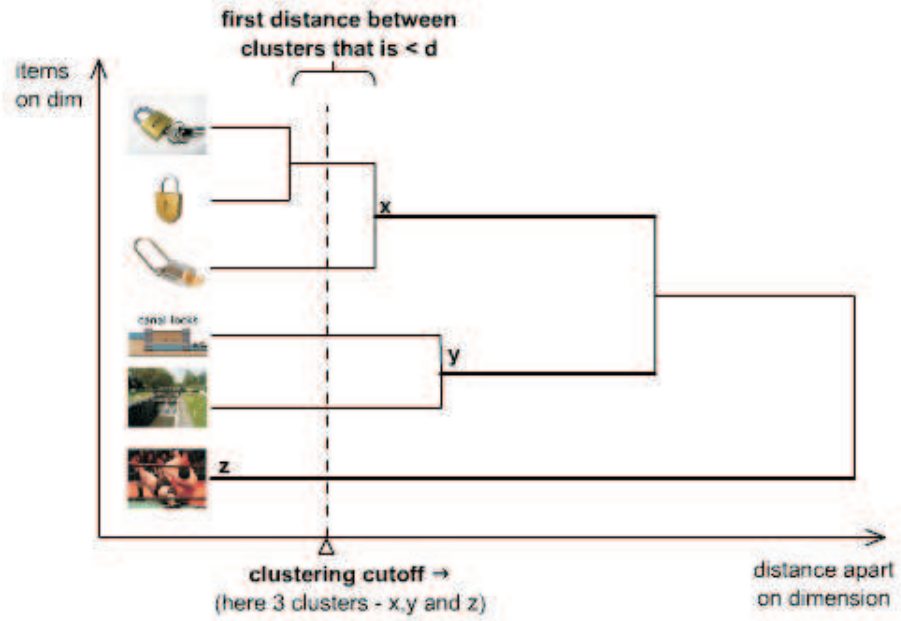

FIGURE 6.1: Determining sense groupings

6.2.1 Lower Mean dwell Time

During trials, it has been observed that the dwell times noted for image searches are significantly lower than those recorded for hypertext document retrieval. This is entirely logical. In general, a user can determine the relevance or non-relevance of a single picture much faster that they can perform the same evaluation on a hypertext document (which may contain any number of paragraphs, headings, lists, images etc.). In any respect, these lower dwell time values do not affect the convergence algorithm deleteriously as we use a mean dwell time *specific to the media environment* to weight them.

6.2.2 Presenting Results

Having completed the steps above and having populated the feature space to a sufficient degree it is now possible to use this stored information to auto-categorise any list of results supplied by the Google Image search engine (i.e. SENSAI can now switch

to *active* mode). When the search coordinator script passes a list of image results for query term k to the categorisation module, SENSAI begins by clustering D_k and determining the correct number of sense groupings for that dimension, as explained in §4.4.1 and illustrated in FIGURE 6.1. SENSAI then filters the Google-Image results through these groupings, so that each image is assigned to a sense category. Images not recorded in the feature space at this point are assigned to a catch-all *unclassified* category.

The user is subsequently presented with *sets* of images that may satisfy the submitted query, with each image group corresponding to a different sense of the query term. Each set is represented as a whole by a centroid image, which serves as a semantic signpost or key to guide the user. The user can easily move between different sets of thumbnails (and therefore different senses of the query term) by clicking on the associated signpost image[2]. An annotated screen-shot of a set of SENSAI image search results illustrating these various features is shown in FIGURE 6.2.

6.2.3 Determining Meaning

As mentioned above, SENSAI enables the resolution of meaning at both ends of the searching operation. Firstly, any ambiguity in the meaning of the query is resolved by presenting the user with the various senses that query could carry and then requiring them to make an affirmative choice. Viewed in this way, *the user* is the sense-resolution system and SENSAI is merely the mechanism through which sense alternatives (gathered from other users) can be brought into focus.

Secondly, and at the other end of the searching process, SENSAI then resolves the *meaning* of each image in the searched collection. Meaning is discovered by examining the co-occurrence of attributes (representing images) in the feature space. If image x co-occurs with a cluster of images about y, then through association it can be assumed that x is also about y. Needless to say, this derived meaning is peculiar to the users of the system and of course subject to change.

[2]Although we experimented with the idea of using a composite centroid image that abstracted details from all of the images it was intended to represent, this technique proved troublesome and is perhaps better suited to text search [266].

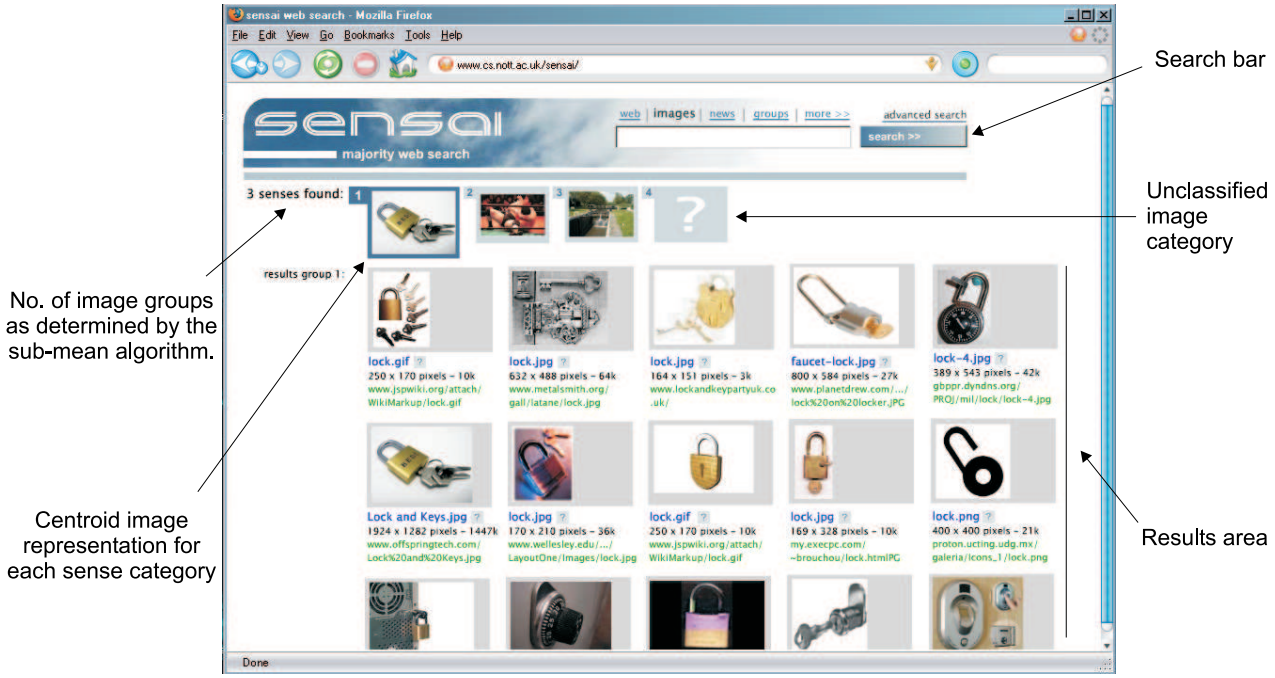

FIGURE 6.2: Categorised search results

6.3 CBIR

The discussion above relates to a certain type of image retrieval, where text-based queries are submitted to the search system and the images themselves are indexed using text-based IR techniques [234]. However, a second and perhaps more powerful technique, known as content based image retrieval (CBIR)[218], may also benefit from co-active search techniques. CBIR addresses a collection of images on a visual or perceptual level [321]. Each document in the collection is examined and low-level features (for example, colour, shape, regions, texture, intensity, spatial characteristics etc.) are extracted and stored [298, 60, 111, 294, 285]. The user of a CBIR system is then invited to provided a sketch or representative image as the query *instead* of entering keywords, a type of query by example [39, 99]. Low level features are extracted from this pictorial query and compared with the features extracted from the searched collection. Those images satisfying some similarity criterion with respect to the query are then returned to the user as results.

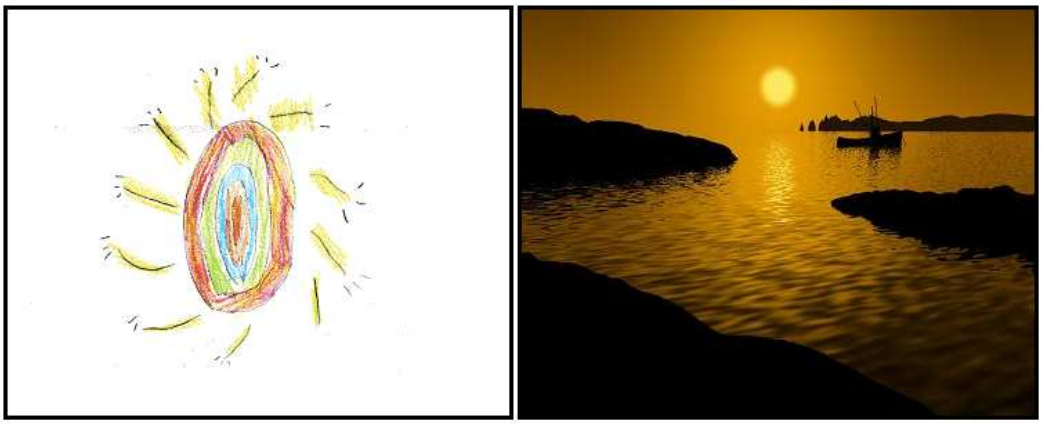

FIGURE 6.3: The Semantic Gap

One of the discernible problems in CBIR is known as the *semantic gap*, being the discrepancy between '..the relatively limited descriptive power of low level imagery features and the richness of user semantics' [61] (see also [280]). This problem is most apparent when a set of results contains numerous images that are similar to the query image in terms of feature similarities, but very different in terms of high-level concepts (as interpreted by the user). To illustrate the semantic gap, let us assume that a user provides a CBIR system with a symbol or pictogram in lieu of a literal graphical representation, an image which indicates a concept in semi-abstract terms rather than directly mapping to some image where that concept is embodied[3] The sketch the user submits is an approximate circle surrounded by radiating lines, a symbol commonly agreed to represent the Sun despite the obvious differences between any realistic depiction of that object and the symbol itself (see FIGURE 6.3). A standard CBIR engine is likely to falter when it encounters a query of this sort because it will inevitably try to find candidate images that resemble it in some way, rather than penetrating the underlying semantic meaning.

[3]Symbols are present from our earliest years of our education [190, 69]. It is not unreasonable to assume that a user of a CBIR system who is unaccustomed to illustrating may revert to the use of symbols to express their information need.

6.3.1 A Co-active CBIR System?

It is hypothesised that a CBIR system implementing the co-active techniques discussed in this thesis would suffer no such semantic gap. Assuming a feature space in which one dimension can represent *either* a query term *or* a pictorial query, and furthermore assuming the means to generalise pictorial queries so that similar sketches can be grouped together, the SENSAI system could comfortably support content-based indexing.

Therefore, in the immediate future, a co-active image retrieval system implementing content-based indexing alongside word-based indexing will be produced. Hopefully, this dual mode system will combine the strengths of both approaches to improve retrieval effectiveness and user satisfaction [194].

6.4 Related Work

Image search engines that assemble the results of a given query into similar categories have been discussed in the literature [189, 253], as have retrieval systems employing cluster analysis of some feature space in order to determine appropriate search results [52, 165, 279]. The crucial difference between a co-active search system and any other contemporary image retrieval engine is that the data used to categorise the image collection is the product of a massive relevance feedback exercise. Whereas other image retrieval systems solicit relevance feedback on a per-person level to improve search results [195, 205, 306], with any adaptation that the system can offer affecting only that particular user, SENSAI agglomerates the workings and careful decisions of an *entire user population* into fluid image classification data. The only real analogue of this general approach was developed by von Ahn & Dabbish, who designed a system which addressed the image problem through entertainment, encouraging people to 'label images whilst enjoying themselves'. [308].

6.5 Conclusion

In this chapter the minor modifications necessary to adapt co-active techniques to a new search environment have been demonstrated. Having described one conversion of the SENSAI system it becomes possible to speculate on others. The field of information retrieval as a whole encompasses the search and selection of a wide spectrum of media types in addition to those discussed in CHAPTERS 5 - 6 (for example, moving imagery, spoken word documents, musical audio data etc. [115, 310, 283]). Assuming the presence of some underlying search engine that 'understands' a particular media type and a sufficiently large user population, there is no reason why the techniques discussed in this thesis should not be applied to each one in turn.

Chapter 7

Further Work and Conclusion

Co-active search is an elegant yet simple solution to a complex problem, and in its single-term core format offers a powerful tool for semantic modelling and the resolution of ambiguity. However, the *real potential* of co-active search lies in the fact that its simple approach encourages powerful and surprisingly flexible extensions. In this chapter several of these extensions - which will form the basis of future work - are discussed. One of these extensions is the use of a fully populated feature space alongside an AHG system, bringing us full circle to the opening chapters of this thesis.

7.1 Further Work

7.1.1 Multi-term Queries

Experiments thus far have concentrated on single word queries and the obvious extension to this work is consideration of queries consisting of more than one term. Queries composed of more than one word are significantly less likely to be ambiguous than some single term query expressing some part of the same idea. This is because longer queries are contextually richer. For example, the word *field* in isolation is highly ambiguous, presenting at least 9 separate senses. However, when the query is expanded to *field + cricket*, the number of applicable senses is dramatically reduced.

Of course, this is not to say that there is no ambiguity in multiple word queries. In fact, there exists a specific problem class known as *lexical phrases* [174]. Lexical phrases are typically composed of two to three words, and are often proper nouns

or technical concepts. For example, consider the phrase *back end*. Used in relation to some temporal duration this phrase has a certain specific meaning (i.e. *the back end of a sports season*). However, used with respect to to the architecture of some computer science project the phrase has a quite different meaning (i.e. *the back end of the system was an SQL server*).

The process by which co-active intelligence can be used to disambiguate these two very different meanings is almost identical to the process used for single term queries, with the exception that here both D_{back} and D_{end} must be considered simultaneously. To accomplish this task one dimension is projected over the other, and the resulting two dimensional space is analysed to determine the viable sense clusters. Interestingly, this process reveals an incredibly valuable facet of co-active intelligence - projecting one dimension over another reveals the *emergent semantics* of that particular space. To illustrate, let us assume that some hypothetical group of users have been searching an information space using the SENSAI system. Let us also assume this group of users have been submitting *single term queries* exclusively, and that the particular queries *back* and *end* have been passed to SENSAI a number of times in the past. Accordingly, the dimensions D_{back} and D_{end} are well populated, as shown in FIGURE 7.1.

Now let us assume that a user enters the query *back + end*, and that this is the first time SENSAI has received such a query. By projecting D_{back} over D_{end} and considering the result SENSAI can identify viable sense groupings for this query despite the fact it has *never been searched for before*. As illustrated in FIGURE 7.1, many of the search artifacts present in this two dimensional space will have only been associated with one of the the search terms. As a result the two dimensional space is heavily populated about the origins. However, some of the search objects will have been associated with both search terms, and it is by clustering these sub-vectors that the sense groupings for this particular query can be determined.

This process of semantic discovery through projection has the potential to uncover many interesting and unexpected associations. This is a particularly exciting property of co-active intelligence and merits further serious research.

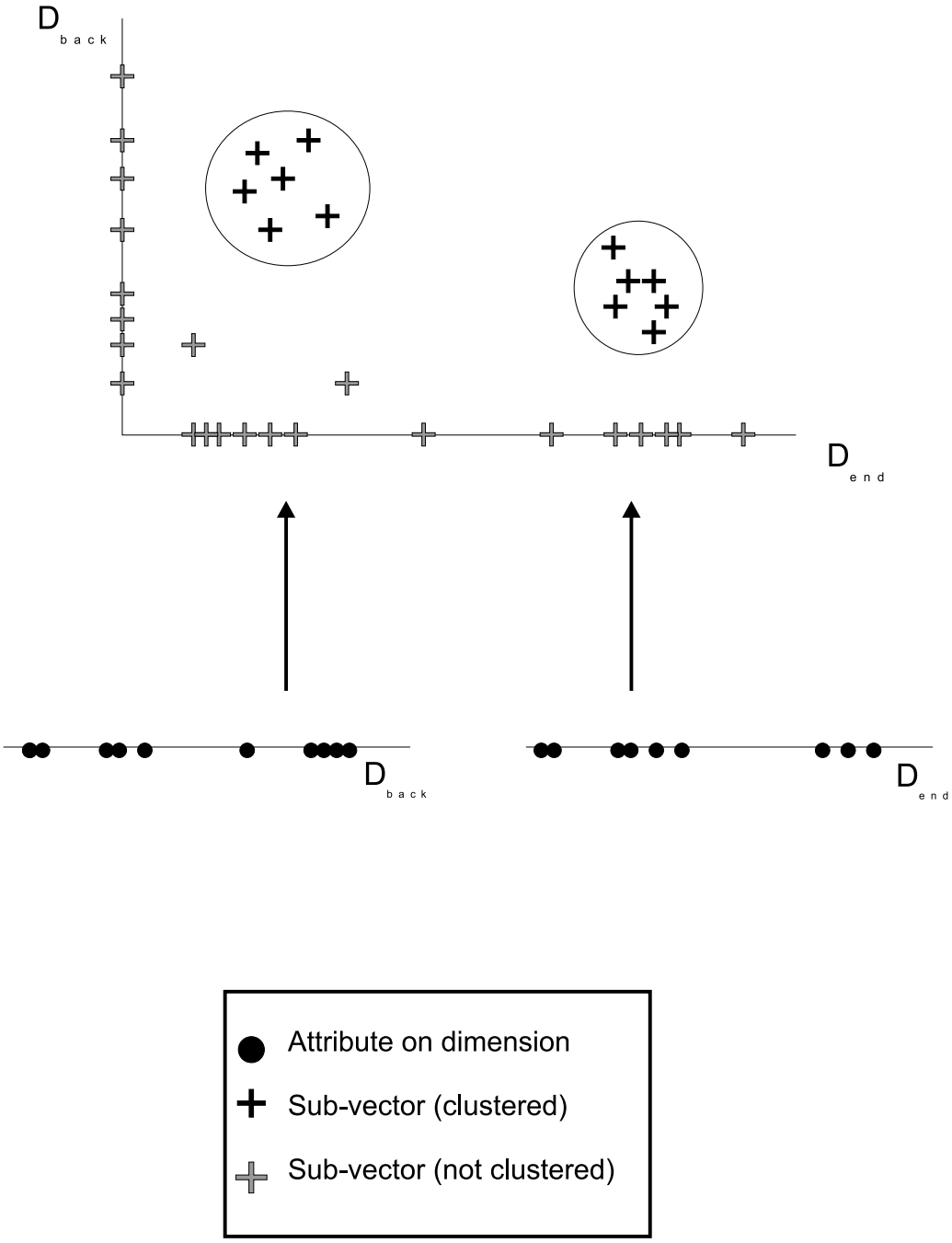

FIGURE 7.1: Sense categorisation in two dimensions

FIGURE 7.2: A Simple Visualisation Tool for SENSAI

7.1.2 Visualisation

The current method of presenting the search results to the user is intended to emulate conventional search engines, but a wide variety of alternative visualisations are possible and indeed encouraged. During the initial text and image experiments a primitive visualisation tool was developed that modelled single dimensions as a series of interconnected blocks, the colour of each block dependent upon the frequency of attributes in that portion of the dimension that the block represented. This visualisation was implemented as a web page with live calls to the SENSAI database, and provided a very simple mechanism for observing and analysing the formation of clusters during user testing. A screen-shot of the web tool is shown in FIGURE 7.2.

Visualisations that address more than one dimension at a time are particularly interesting. Where a query is made up of more than one single term a two- or three-

dimensional representation of the semantics of the search results could be constructed. This image map might enable users to perform interactive operations on the visualised clusters, allowing them to refine previous searches by splitting clusters or broaden out a stalled search by relaxing the clustering criterion. Any work in this direction would be heavily guided by other authors who have previously discussed the visualisation of IR search results [9, 80, 299, 274, 24, 145]. A further source of inspiration might be found in any one of the successful visualisation software packages currently available (for example, [239]).

7.1.3 Open-Tagging

Development of co-active theory and practice has tellingly synchronised with a wider trend towards the decentralisation of resource descriptions. An Internet site which enables users to upload and annotate their image collections with content describing notes and conceptual tags has recently caught the attention of both the computer press and the blogging community [2][1]. The fundamental difference between the functionality offered by this particular site and many others of a similar ilk is that users are free to annotate the images of *other users* in addition to their own material. The wholesale success of this approach has affirmed that an open mandate to tag other people's data is succeeding in the face of initial fears concerning the possibility of malicious or disruptive activity.

The use of open tagging protocols alongside co-active inspired retrieval systems is eagerly anticipated. One of the difficulties experienced with the SENSAI system concerns the high volume of search objects the system initially returns as *unclassified*. An unclassified object is an image or document suggested by Google as relevant to the user query which does not already appear in the feature space. The current procedure when such an object is passed to the categorisation module is to place that object in a catch-all category having created a new (implicit) feature vector to represent it. Unfortunately, it has been observed in our earliest tests that where several strong sense groupings are available to the user the unclassified results group

[1]For a full text equivalent, which allows users to tag a web site using the bookmark function of their preferred browser, see [270]

is for the most part ignored. This of course raises serious issues with regard to the convergence algorithm. If an image or a document is initially placed into the low visibility unclassified group, it may never be the subject of enough associations to move it into the correct sense grouping. Relegated in this way, the resource will have effectively disappeared from the search space.

However, by incorporating open tagging protocols into a co-active system this problem can be addressed directly. By interleaving objects from the unclassified group into the primary sense groupings, employing user provided open tags as a prediction of compatibility in that grouping, the quality of associative data available to the system is greatly increased. The simultaneous and independent development of open tagging protocols and co-active techniques for information retrieval is a rare and auspicious coincidence which demands additional consideration.

7.1.4 Active/Passive Modes

SENSAI is a two-state system. In *passive* mode it acts as a simple information conduit, passing queries from the user to the underlying search engine, then routing the results back in the opposite direction. In *active* mode, SENSAI uses the navigational data collected whilst in *passive* mode to provide a sense categorisation service. As discussed in CHAPTERS 4- 6, this move from one operational state to another presently affects *all* dimensions uniformly.

Whilst this configuration is appropriate for small testing runs consisting of no more than 10 queries, there are serious doubts about its suitability as part of a full scale deployment. When SENSAI provides its categorisation service to a large search engine, the influx of unseen queries - which affects the demand for new dimensions - will be unchecked. Furthermore, certain queries will occur much more frequently than others. The net result of exposure to a live user base is therefore likely to be a patchy, uneven population of the feature space - at any given time, some of the dimensions would be ripe for clustering whilst others would consist of little more than noise.

Given the above, an operational move from one processing state to another that affects every dimension in the feature space is untenable. Instead, there is a real need

to determine the most appropriate mode for the system on a per-query basis, with the result dependent upon the *cluster tendency* of the dimension(s) involved [91, 62]. This would work in the following way. Having received query term k, SENSAI would examine dimension k and estimate the validity of the outcome of a clustering operation prior to it being performed. Where the cluster tendency is good, SENSAI would serve the results in *active* mode. Where the cluster tendency is poor, SENSAI would serve the results in *passive* mode and continue observing. Future work will certainly include the development of a per-dimension cluster tendency algorithm along these lines, with the precise metric for evaluating the suitability of clustering to be determined at some later date.

7.1.5 Potential Exploits

At present, the SENSAI system is vulnerable to a category of exploits we have labelled mischievous associations. A mischievous association is made when a user deliberately manipulates the system to associate some resource with a query term for financial gain, to make some type of public statement, or to mis-lead, shock or amuse others. The implementation of negative feedback (see §5.3.2) and dwell time weights will serve to curb this abuse to some extent, but high frequency automated attacks are capable of creating incorrect associations faster than they can be removed by responsible users. Therefore, the future is likely to prompt the investigation of anti-flooding techniques - as practised on news groups and forums - to provide an additional measure of protection.

7.1.6 Ostensive Associations

Information spaces like the Internet are extremely volatile [36, 64, 97, 63, 104]. The content of a hypertext document that formed part of a recorded association can change over time, so that the associations to which it belonged - though valid at the time - should no longer apply. In §3.8.4, the practice of ostensive relevance feedback was discussed, whereby relevance judgements received from the users of a search system are weighted so that more recent judgements have greater impact than their

earlier counterparts [55]. With little effort, a system of ostensive associations can be implemented for any co-active search system, so that the impact of an association on the feature space of that system deteriorates in a similar fashion.

One simple implementation of time-weighted associations could utilise the Unix timestamp. Periodically, each recorded association would be re-weighted using the following formula:

$$Weight\ of\ Association\ x\ =\ 1 - \left(\frac{TN - TS_x}{SysStart}\right) \tag{7.1}$$

where

TN expresses the current Unix time in seconds

TS_x is the timestamp of association x

$SysStart$ is the first timestamp recorded by the system.

TABLE 7.1 shows a worked example, assuming a current Unix time of 1119528000 (23rd June 2005, 12:00:00) and that the system became operation at 1114257600 (23rd April, 2005, 12:00:00).

TABLE 7.1: Ostensive Weightings

Association Recorded	Timestamp	Ostensive Weighting
Same Day	1119528000	1
One Day Earlier	1119441600	0.983606557
One Week Earlier	1118923200	0.885245902
One Month Earlier	1116849600	0.491803279
Two Months Earlier	1114257600	0

Further work on the viability and implications of ostensive associations within a feature space should be pursued at some later date.

7.1.7 Regionalisation

Another factor to be considered in the future is the extent to which search results served to a particular user should be modified according to their location. Alternative meanings attached to the same word in different countries (or regions of the same

country) present yet another aspect of the word sense problem[2]. Further work is necessary to investigate the viability of a region-specific component of the SENSAI system, and the effect that such a component would have on the underlying semantic modelling techniques.

7.1.8 Associative Indexing

As discussed in CHAPTER 6, image retrieval engines commonly use text-based IR techniques to extract terms describing an image from a parent document (e.g. a web page). Associative indexing is a refinement to this process made possible by the unique features of the SENSAI system. In this procedure, textual analysis is performed on the parent document of the target image *and* the parent documents of all images co-occurring with the target image on a specified dimension. This simple statistical operation, which could use the text processing techniques described in §5.2.4, has the potential to provide descriptive key words superior in quality to those obtained using conventional word-based indexing [234].

7.1.9 Transitive Association

One further area that remains to be studied is the possibility of *transitive associations*. Transitive associations can be drawn from a set of session records so that a single cluster created by the user population is associated with another cluster representing a similar concept, albeit specified with an alternate syntax. For example, the English word 'football' will generate a cluster of attributes representing text documents concerning that subject. Likewise, the American word 'soccer', which describes the same activity, will also generate a cluster of attributes, representing (presumably) very similar documents. Transitive association would allow us to automatically recognise these synonymous clusters so that a query search for 'football' would return 'soccer' results automatically, irrespective of whether a 'bridging document' (which uses both terms) is part of the collection. Further inquiry into the mechanisms through which inter-cluster similarities can be identified and exploited in an automatic fashion will

[2]For example, the word *chips* means quite a different thing in America than it does in the UK

be forthcoming.

7.1.10 A Co-active AHG?

Finally, another possible application for co-active theory lies in the development of a automatic hypertext generation system that is responsive to word meaning. This could be implemented by developing simple algorithms capable of suggesting links between a set of documents and pairing these algorithms with a well populated feature space. This feature space would be generated by the users of a otherwise unrelated search system (as discussed in chapter 5) and would enable the link creation service to make sense-aware suggestions. Pleasingly, this hypothetical but quite realisable extension completes a circular journey begun all the way back in chapter 1 with the LIAR engine.'

7.2 Conclusion

In this thesis the theory of co-active search has been introduced, a democratic and flexible solution to the Gordian problem of lexical ambiguity in query-based search. One of the key advantages of this technique is that it does not rely on any external source of manually compiled information to perform word disambiguation - instead of dictionaries, theasauri or tagged text it utilises the discriminatory power of a transparently polled user base to quickly determine the possible meanings a word can carry. This automated procedure neatly avoids the necessity of an extended knowledge capture process followed by interminable amendments.

Also discussed was a radical philosophy that engages with the users of a system as a collective intelligence capable of resolving problems quite intractable to automated processing [297]. The challenge now lies in identifying other areas in which the use of co-active strategies may be successful. However, the lesson from this work is clear. The experience and wisdom of the users as they interact with any system is a vital and unappreciated resource. The decisions users make and the decisions they forego; the inter-relationships they identify and the assumed relationships they disprove - these are all valuable commodities, available at very little cost. Any initiative centred on

the cultivation and development of this resource is to be encouraged as a matter of priority.

Appendix A

Stop List

TABLE A.1: Fox's Stop List for General Text

a	about	above	across	after	again	against	all
almost	alone	along	already	also	although	always	among
an	and	another	any	anybody	anyone	anything	anywhere
are	area	areas	around	as	ask	asked	asking
asks	at	away	back	backed	backing	backs	be
because	become	becomes	became	been	before	began	behind
being	beings	best	better	between	big	both	but
by	came	can	cannot	case	cases	certain	certainly
clear	clearly	come	could	did	differ	different	differently
do	does	done	down	downed	downing	downs	during
each	early	either	end	ended	ending	ends	enough
even	evenly	ever	every	everybody	everyone	everything	everywhere
face	faces	fact	facts	far	felt	few	find
finds	first	for	four	from	full	fully	further
furthered	furthering	furthers	gave	general	generally	get	gets
give	given	gives	go	going	good	goods	got
great	greater	greatest	group	grouped	grouping	groups	had
has	have	having	he	her	herself	here	high
higher	highest	him	himself	his	how	however	if
important	in	interest	interested	interesting	interests	into	is
it	its	itself	just	keep	keeps	kind	knew
know	known	knows	large	largely	last	later	latest
least	less	let	lets	like	likely	long	longer
longest	made	make	making	man	many	may	me
member	members	men	might	more	most	mostly	mr
mrs	much	must	my	myself	necessary	need	needed
needing	needs	never	new	newer	newest	next	no
non	not	nobody	noone	nothing	now	nowhere	number
numbered	numbering	numbers	of	off	often	old	older
oldest	on	once	one	only	open	opened	opening
opens	or	order	ordered	ordering	orders	other	others
our	out	over	part	parted	parting	parts	per
perhaps	place	places	point	pointed	pointing	points	possible
present	presented	presenting	presents	problem	problems	put	puts
quite	rather	really	right	room	rooms	said	same
saw	say	says	second	seconds	see	sees	seem
seemed	seeming	seems	several	shall	she	should	show
showed	showing	shows	side	sides	since	small	smaller
smallest	so	some	somebody	someone	something	somewhere	state
states	still	such	sure	take	taken	than	that
the	their	them	then	there	therefore	these	they
thing	things	think	thinks	this	those	though	thought
thoughts	three	through	thus	to	today	together	too
took	toward	turn	turned	turning	turns	two	under
until	up	upon	us	use	uses	used	very
want	wanted	wanting	wants	was	way	ways	we
well	wells	went	were	what	when	where	whether
which	while	who	whole	whose	why	will	with
within	without	work	worked	working	works	would	year
years	yet	you	young	younger	youngest	your	yours

Appendix B

Search Agendas

TABLE B.1: Agendas - Data Set One (pt. 1)

Query Term	Agenda
apple	1. You are compiling a report on the history of Apple Computers. Find three relevant web pages about the topic. 2. You are writing a web site about the cultivation and care of apple trees. Find three web sites you might like to link to. 3. You are travelling to New York (the 'Big Apple') this Autumn. Find three web pages you might find useful.
flush	1. You are playing Texas Hold'em poker later this week. Find three useful web sites giving information on the chances of building a flush. 2. Your toilet handle is broken. Find three websites that could help you repair the flush. 3. Your child is unwell and is having hot flushes. Find three web pages you might read for advice.
skate	1. Your friend, a keen fisherman, wants information about the various rays of the genus raja (skates). Find three web sites you can recommend to him. 2. Find three web sites giving information on skate-boarding tricks. 3. Your child has just taken up in-line skating and you are concerned for their safety. Find three helpful web sites about the subject
tube	1. You are writing a report on the history of the London underground. Find three good web pages about the topic. 2. You want to send a valuable poster through the post. Find three web sites selling cardboard tubes that could be used to protect it. 3. Find three web sites with pictures of surfers 'in the tube'.
bat	1. Find three web sites giving information about bats, the flying mammals. 2. You want to buy a baseball bat. Find three retailers in the UK. 3. You want information about the construction of a cricket bat. Find three useful web sites.

TABLE B.2: Agendas - Data Set One (pt. 2)

Query Term	Agenda
ball	1. You are experiencing pain in the ball of your foot. Find three web sites for advice. 2. You are curious about how they put golf balls together. Find three web sites on their construction. 3. Your daughter will soon attend a debutante ball. Find three sites with more information for her.
fence	1. You are building a fence in your garden. Find three informative web sites. 2. You want to join a fencing club in the UK. Find three useful web pages. 3. The term 'fence' is slang for someone who handles stolen goods. Find three dictionary definitions.
bowling	1. You have just taken up cricket. Find three web sites with tips for better bowling. 2. You have just taken up 10 pin bowling. Find three web sites that can tell you the rules. 3. You want to take up Crown Green bowling. Find three websites about UK clubs.
spirits	1. You have become a Christmas grouch. Find three web sites that can help you get back into the Christmas Spirit. 2. You want to distill your own spirits. Find three websites with advice and equipment for sale. 3. You are interested in the occult. Find three web sites on the ghosts and spirits.
lock	1. You need to fit an interior door lock. Find three web sites with advice. 2. You are learning to wrestle. Find three web pages on the various locks and hold a wrestler uses. 3. You want to know what a lock in a canal does, and why. Find three web pages about this topic.

TABLE B.3: Agendas - Data Set Two (pt. 1)

Query Term	Agenda
port	1. You need to know more about computer ports. Find three web pages that provide good information about the subject. 2. You want to know how the alcoholic beverage port is produced. Find three informative web sites. 3. You need more information on shipping ports in the UK. Find three useful web pages.
tick	1. Your dog has picked up a tick whilst walking in the park. Find three web pages describing how to remove the parasite. 2. A 'tick' is a short term technical indicator used in economics. Find three web pages with dictionary definitions of this particular use of the word. 3. Your nephew wants to buy a comic book. The lead character is called the 'Tick'. Find three web pages that sell it.
bug	1. Find three pages giving you more information on the River Bug. 2. Find three good definitions of a computer bug. 3. You suspect your partner is cheating on you. Find three places to buy listening devices (bugs). 4. You own a classic VW Beetle (or 'bug'). Find three sites describing car clubs, rallies or festivals for that particular car.
fly	1. You have taken up sewing and you would like to sew a fly onto a garment. Find three web pages that can help you. 2. You need facts about animals of the order diptera (flies). Find three web pages with useful information. 3. Your mother is a keen fisherwoman. Find thee UK stores that sell fishing flies.
net	1. You are writing an essay on boat fishing techniques involving nets. Find three useful web sites. 2. You need a dictionary definition for 'net' as in net profit. Find three web sites that can provide you with one. 3. You are writing a history of the Internet. Find three informative web sites

TABLE B.4: Agendas - Data Set Two (pt. 2)

Query Term	Agenda
memory	1. You want to improve your memory. Find three useful web sites. 2. You want to buy some memory for your computer. Find three web sites that can help you. 3. Find out more about the term 'memory' in relation to immunology. Find three web sites with useful information.
chips	1. You want to find the best fish and chip shop in the UK. Find 3 web pages that might help. 2. You are interested in buying a new processor chip for your PC. Find three web pages that sell them. 3. You want to buy some clay poker chips for your regular game. Find three online stores.
pen	1. You want to buy a nice new writing pen. Find three online stores. 2. You want more information about female swans (pens). Find three helpful sites. 3. You are thinking about keeping animals in your backyard. Find some web sites about animal pens.
paris	1. You will soon travel to Paris, France. Find three web sites with useful information. 2. You will soon travel to Paris, Texas. Find three web sites with useful information. 3. You need to buy some plaster of paris. Find three web stores that can sell it to you.
cavalier	1. You are thinking about buying a used Vauxhall Cavalier. Find three web sites that may help you. 2. Someone has just accused you of having a 'cavalier' attitude. Find three definitions of this word. 3. You would like to buy a Cavalier King Charles spaniel. Find three web sites that might help you. 4. You are writing an essay on the Cavalier army that supported King Charles I. Find three web pages that might be of interest.

Appendix C

System Log

```
SESSION (046b5e07ebddc1570bcfd4cbbb7e512a)
bug association:  www.bugjam.co.uk/ [0.63480592491378] [4]
association:  uk.bizrate.com/buy/products-cat- id--14010200,keyword--Listening Devices.html [0.175271552] [3]
association:  www.ozspy.com.au/audio/audio.asp [0.1779058624] [3]
association:  poland.pl/nature/regions/niziny-srp/bug/description.htm [0.6574] [1]

tick association:  www.lyme.org/ticks/removal.html [0.60798384] [1]
association:  www.amazon.com/exec/obidos/ tg/detail/-/B0000AUHQE?v=glance [0.6078328] [4]
association:  www.anapsid.org/lyme/howtoremoveticks.pdf [0.580751328] [1]

SESSION (28e4e5a8a974892537eca2ae3e15dfa5)
bug association:  bugclub.org/beginners/software/bugs.html [0.61030676309333] [2]
association:  onlinedictionary.datasegment.com/word/bug [0.62881422717583] [2]
association:  encyclopedia.laborlawtalk.com/Computer-bug [0.59236995043983] [2]
association:  www.stevenagevwclub.co.uk/links.html [0.65887082091378] [4]
association:  www.cvwoc.co.uk/pages/links.htm [0.63480592491378] [4]
association:  www.ccdata.com/browse/Owners-Marques-V-Volkswagen [0.64321888] [4]
association:  www.bugjam.co.uk/ [0.63480592491378] [4]

memory association:  www.sciencedaily.com/releases/2003/11/031112073642.htm [0.22257277333333] [3]
association:  www.medicalnewstoday.com/medicalnews.php?newsid=15950 [0.1649] [3]
association:  www.cat.cc.md.us/courses/bio141/lecguide/ unit3/humoral/antibodies/memory/memory.html [0.29714514143492] [3]
association:  www.emoryhealthcare.org/press-room/ ehc-news/2004/Nov/Long-term-Immune-Memory-Cells.html [0.28922962427937] [3]

net association:  www.investordictionary.com/definition/Net profit.aspx [0.63749314285714] [2]
association:  www.hll.com/HLL/investors/glossary.html [0.63822335238095] [2]
association:  www.wordwebonline.com/en/NETPROFIT [0.63822335238095] [2]
association:  www.nethistory.info/ [0.6505] [3]
association:  www.internetvalley.com/intval.html [0.6505] [3]
association:  www.davesite.com/webstation/net-history.shtml [0.63680857142857] [3]
association:  www.walthowe.com/navnet/history.html [0.63680857142857] [3]
association:  www.isoc.org/internet/history/ [0.63680857142857] [3]

paris association:  www.artlex.com/ArtLex/p/plaster.html [0.4396] [3]
association:  www.pharmacy-online.ca/quick- search.jsp?criteria=PLASTER OF PARIS [0.4396] 3
association:  familycrafts.about.com/ cs/miscjewelry/a/blplastercast.htm [0.43749866666667] [3]
association:  www.misterart.com/store/view/ 003/group-id/965/DAP-Plaster-of-Paris.htm [0.42489066666667] [3]
association:  www.hugahorse.org/DA/ DAP--white-plaster-of-Paris--dry-mix---4-lbs..html [0.42489066666667] [3]
association:  www.theparisnews.com/ [0.43949493333333] [2]
association:  www.cityofparistx.com/ [0.43949493333333] [2]
association:  lone-star.net/mall/txtrails/paris.htm [0.43780686222222] [2]

pen association:  www.websterspenshop.co.uk/ [0.4886] [1]
```

```
association:  www.euroffice.co.uk/itm-groups.asp?SBCAT=267 [0.4886] [1]
association:  uk.bizrate.com/buy/products- -cat-id--212,keyword--Pen Refills.html [0.4886] [1]
association:  www.needapresent.com/shop/ get-ProductDetail.asp?PID=1155 [0.4886] [1]
association:  buy.ebay.co.uk/parker [0.4886] [1]

port association:  www.ctips.com/rs232.html [0.4401] [1]
association:  www.abilityhub.com/information/ports.htm [0.4401] [1]
association:  www.computerhope.com/network/ports.htm [0.4401] [1]
association:  computer.howstuffworks.com/usb.htm [0.4401] [1]

tick association:  www.answers.com/topic/uptick [0.6775] [2]
association:  www.uptick.com/ [0.6775] [4]
association:  www.investordictionary.com/definition/Tick.aspx [0.676792] [2]
association:  tick Definition [0.6775] [2]
association:  www.specialinvestor.com/terms/947.html [0.676792] [2]
association:  www.answers.com/topic/tick [0.668296] [2]

SESSION (2a6b9f901e503b137e9326c1fd4608bc)
bug association:  en.wikipedia.org/wiki/Computer-bug [0.45725777905635] [2]
association:  www.ozspy.com.au/audio/audio.asp [0.1779058624] [3]
association:  uk.bizrate.com/buy/products--cat- id--14010200,keyword--Listening Devices.html [0.175271552] [3]
association:  www.edirectory.co.uk/pf/pages/moreinfoa.  asp?pe=FFDIJHQ- Bug em Listening Device [0.1673686208] [3]

cavalier association:  www.wordwebonline.com/en/CAVALIER [0.527] [2]
association:  www.answers.com/topic/cavalier [0.527] [2]
association:  onlinedictionary.datasegment.com/word/cavalier [0.527] [2]
association:  dict.die.net/cavalier/ [0.527] [2]

chips association:  www.reviewcentre.com/reviews14348.html [0.2252056] [1]
association:  www.restaurantspy.com/ uk/england/kent/no1fishandchips.htm [0.259228] [1]
association:  www.harryramsdens.co.uk/about/downloads/statistics.doc [0.2195352] [1]
association:  www.ukwebpages.co.uk/bardsleys/web-wall.htm [0.2195352] [1]
association:  www.thefoodplace.co.uk/restaurants/ 19876/Best Fish and Chips in Waltham Cross/ [0.266316] [1]

fly association:  www.diynet.com/diy/na-serging/ article/0,2025,DIY-14143-2277030,00.html [0.5424] [1]
association:  www.hgtv.com/hgtv/cr-sewing-buttons-zippers/ article/0,1789,HGTV-3325-1377114,00.html [0.5424] [1]
association:  sewing.patternreview.com/ cgi-bin/review/readreview.pl?ID=623 [0.5424] [4]
association:  sewing.about.com/library/sewnews/qa/aaqa0500b.htm [0.5424] [1]
association:  www.sewnews.com/library/sewnews/qa/aaqa0702c.htm [0.5424] [1]
association:  www.sewing.org/enthusiast/html/el-flyfrontzipper.html [0.51637333333333] [1]

memory association:  www.memorex.com/ [0.6504336] [5]
association:  www.laptopshop.co.uk/laptop-memory.htm [0.63963422943492] [2]
association:  www.memoryx.net/memory.html [0.64348242143492] [2]
association:  Orca.co.uk [0.68272596276825] [2]
association:  www.computermemoryupgrade.net/ [0.64348242143492] [2]
association:  Crucial.com/uk/ [0.7286] [2]

net association:  www.wordwebonline.com/en/NETPROFIT [0.63822335238095] [2]
association:  www.hll.com/HLL/investors/glossary.html [0.63822335238095] [2]
association:  dict.die.net/profit/ [0.6505] [2]
association:  www.investordictionary.com/definition/Net profit.aspx [0.63749314285714] [2]
association:  www.investorwords.com/3259/net-profit.html [0.64698586666667] [2]

paris association:  www.guidetocinema.com/paris.html [0.42763640888889] [4]
association:  www.city-data.com/city/Paris-Texas.html [0.43780686222222] [2]
association:  lone-star.net/mall/txtrails/paris.htm [0.43780686222222] [2]
association:  www.cityofparistx.com/ [0.43949493333333] [2]
association:  www.theparisnews.com/ [0.43949493333333] [2]

pen association:  buy.ebay.co.uk/parker [0.4886] [1]
association:  www.needapresent.com/shop/ get-ProductDetail.asp?PID=1155 [0.4886] [1]
association:  uk.bizrate.com/buy/products- -cat-id--212,keyword--Pen Refills.html [0.4886] [1]
```

```
association:  www.euroffice.co.uk/itm-groups.asp?SBCAT=267 [0.4886] [1]
association:  www.websterspenshop.co.uk/ [0.4886] [1]

port association:  www.epicurious.com/drinking/ wine-dictionary/entry?id=7570 [0.33661013333333] [2]
association:  www.cellartastings.com/en/wine-feature-port-wine.html [0.33661013333333] [2]
association:  www.the-port-man.fsbusiness.co.uk/ [0.32242773333333] [2]
association:  winemaking.jackkeller.net/reques13.asp [0.337792] [4]
association:  www.goodcooking.com/portinfo.htm [0.3389] [2]
association:  www.intowine.com/port.html [0.3389] [2]
association:  homecooking.about.com/library/archive/blwineindex.htm [0.3389] [4]

tick association:  www.anapsid.org/lyme/howtoremoveticks.pdf [0.580751328] [1]
association:  ohioline.osu.edu/hyg-fact/2000/2073.html [0.5843272] [1]
association:  www.choa.org/default.aspx?id=365 [0.5713] [4]
association:  dogs.about.com/cs/disableddogs/qt/tick-removal.htm [0.575857632] [1]

SESSION (31d23823d0834855006191185df50f34)
bug association:  www.riob.org/ag2000/pologne-bug.htm [0.64322208] [5]
association:  poland.pl/nature/regions/niziny-srp/bug/description.htm [0.6574] [1]
association:  encarta.msn.com/Bug-(river).html [0.6574] [1]
association:  www.britannica.com/ebc/article?tocId=9358240 [0.6574] [1]

cavalier association:  www.usedcarmart.co.uk/page-25612.html [0.705] [1]
association:  carsearch.yahoo.co.uk/ctl/do/ drilldown 100354123/Vauxhall,Vauxhall|Cavalier [0.705] [1]
association:  www.carpages.co.uk/info/vauxhall.asp [0.705] [1]
association:  www.cavweb-forums.co.uk/index.php [0.705] [1]

chips association:  www.kelkoo.co.uk/b/a/sbs/uk/gadgets/ keywords/poker chips/type/100138013.html [0.697798] [3]
association:  www.pokerlistings.com/poker-chips [0.697798] [3]
association:  www.gamble.co.uk/delivery.htm [0.697798] [3]
association:  www.pokershopping.co.uk/poker-chips [0.7483] [3]

fly association:  www.diynet.com/diy/na-serging/ article/0,2025,DIY-14143-2277030,00.html [0.5424] [1]
association:  sewing.patternreview.com/ cgi-bin/review/readreview.pl?ID=623 [0.5424] [4]
association:  sewing.about.com/library/sewnews/qa/aaqa0500b.htm [0.5424] [1]
association:  www.sewing.org/enthusiast/html/el-flyfrontzipper.html [0.51637333333333] [1]
association:  www.sewnews.com/library/sewnews/qa/aaqa0702c.htm [0.5424] [1]

memory association:  www.emoryhealthcare.org/press-room/ ehc-news/2004/Nov/Long-term-Immune-Memory-Cells.html [0.28922962427937] [3]
association:  www.cat.cc.md.us/courses/bio141/lecguide/ unit3/humoral/antibodies/memory/memory.html [0.29714514143492] [3]
association:  www.medicalnewstoday.com/medicalnews.php?newsid=15950 [0.1649] [3]
association:  www.sciencedaily.com/releases/2003/11/031112073642.htm [0.22257277333333] [3]

net association:  www.walthowe.com/navnet/history.html [0.63680857142857] [3]
association:  www.davesite.com/webstation/net-history.shtml [0.63680857142857] [3]
association:  www.internetvalley.com/intval.html [0.6505] [3]
association:  www.wordwebonline.com/en/NETPROFIT [0.63822335238095] [2]
association:  www.hll.com/HLL/investors/glossary.html [0.63822335238095] [2]
association:  www.investordictionary.com/definition/Net profit.aspx [0.63749314285714] [2]
association:  www.investorwords.com/3259/net-profit.html [0.64698586666667] [2]

paris association:  www.conciergerie.com/foryou/ attractions/paris/metro-map.html [0.165376] [1]
association:  Website of the city of Paris [0.127552] [1]
association:  www.france.com/ [0.1244] [1]
association:  www.francetourism.com/ [0.127552] [1]

port association:  computer.howstuffworks.com/usb.htm [0.4401] [1]
association:  www.computerhope.com/network/ports.htm [0.4401] [1]
association:  www.abilityhub.com/information/ports.htm [0.4401] [1]
association:  www.ctips.com/rs232.html [0.4401] [1]

tick association:  www.anapsid.org/lyme/howtoremoveticks.pdf [0.580751328] [1]
association:  dogs.about.com/cs/disableddogs/qt/tick-removal.htm [0.575857632] [1]
```

```
association:  www.furlongspetsupply.com/ how-to-remove-ticks-from-a-dog.htm [0.5713] [1]

SESSION (3706a7bf6d39107cf27ebb7be7b76afa)
bug association:  bugclub.org/beginners/software/bugs.html [0.61030676309333] [2]
association:  encyclopedia.laborlawtalk.com/Computer-bug [0.59236995043983] [2]
association:  en.wikipedia.org/wiki/Computer-bug [0.45725777905635] [2]

cavalier association:  www.answers.com/topic/cavalier [0.527] [2]
association:  onlinedictionary.datasegment.com/word/cavalier [0.527] [2]
association:  www.wordwebonline.com/en/CAVALIER [0.527] [2]

chips association:  www.harryramsdens.co.uk/about/downloads/statistics.doc [0.2195352] [1]
association:  www.reviewcentre.com/reviews14348.html [0.2252056] [1]
association:  www.restaurantspy.com/ uk/england/kent/no1fishandchips.htm [0.259228] [1]
association:  www.ukwebpages.co.uk/bardsleys/web-wall.htm [0.2195352] [1]
association:  www.thefoodplace.co.uk/restaurants/ 19876/Best Fish and Chips in Waltham Cross/ [0.266316] [1]

fly association:  www.sewing.org/enthusiast/html/el-flyfrontzipper.html [0.51637333333333] [1]
association:  www.sewnews.com/library/sewnews/qa/aaqa0702c.htm [0.5424] [1]
association:  sewing.about.com/library/sewnews/qa/aaqa0500b.htm [0.5424] [1]
association:  www.hgtv.com/hgtv/cr-sewing-buttons-zippers/ article/0,1789,HGTV-3325-1377114,00.html [0.5424] [1]
association:  www.diynet.com/diy/na-serging/ article/0,2025,DIY-14143-2277030,00.html [0.5424] [1]

net association:  www.walthowe.com/navnet/history.html [0.63680857142857] [3]
association:  www.davesite.com/webstation/net-history.shtml [0.63680857142857] [3]
association:  www.nethistory.info/ [0.6505] [3]
association:  www.internetvalley.com/intval.html [0.6505] [3]
association:  www.isoc.org/internet/history/ [0.63680857142857] [3]
association:  www.investorwords.com/3259/net-profit.html [0.64698586666667] [2]
association:  www.investordictionary.com/definition/Net profit.aspx [0.63749314285714] [2]
association:  dict.die.net/profit/ [0.6505] [2]
association:  www.hll.com/HLL/investors/glossary.html [0.63822335238095] [2]
association:  www.wordwebonline.com/en/NETPROFIT [0.63822335238095] [2]

paris association:  www.theparisnews.com/ [0.43949493333333] [2]
association:  www.cityofparistx.com/ [0.43949493333333] [2]
association:  www.texastwisted.com/attr/eiffeltower/ [0.4396] [2]
association:  www.city-data.com/city/Paris-Texas.html [0.43780686222222] [2]

pen association:  www.websterspenshop.co.uk/ [0.4886] [1]
association:  www.euroffice.co.uk/itm-groups.asp?SBCAT=267 [0.4886] [1]
association:  uk.bizrate.com/buy/products- -cat-id--212,keyword--Pen Refills.html [0.4886] [1]
association:  www.needapresent.com/shop/ get-ProductDetail.asp?PID=1155 [0.4886] [1]
association:  buy.ebay.co.uk/parker [0.4886] [1]

port association:  www.medwayports.com/ [0.1727] [3]
association:  www.doverport.co.uk/ [0.17417733333333] [3]
association:  www.abports.co.uk/southampton/ [0.18304133333333] [3]
association:  www.abports.co.uk/ [0.18304133333333] [3]

tick association:  www.answers.com/topic/uptick [0.6775] [2]
association:  www.uptick.com/ [0.6775] [4]
association:  www.investordictionary.com/definition/Tick.aspx [0.676792] [2]
association:  tick Definition [0.6775] [2]
association:  www.answers.com/topic/tick [0.668296] [2]

SESSION (40ba6a7e89ccfbd559f085dbcebbeec8)
bug association:  www.bugjam.co.uk/ [0.63480592491378] [4]
association:  www.ccdata.com/browse/Owners-Marques-V-Volkswagen [0.64321888] [4]
association:  www.cvwoc.co.uk/pages/links.htm [0.63480592491378] [4]
association:  en.wikipedia.org/wiki/Computer-bug [0.45725777905635] [2]
association:  encyclopedia.laborlawtalk.com/Computer-bug [0.59236995043983] [2]
```

```
association:  onlinedictionary.datasegment.com/word/bug [0.62881422717583] [2]

cavalier association:  www.dogbreedinfo.com/cavalierkingcharlesspaniel.htm [0.527] [4]
association:  www.wordwebonline.com/en/CAVALIER [0.527] [2]
association:  www.cavaliers.co.uk/ [0.527] [4]
association:  www.cavalier.on.ca/ [0.527] [6]
association:  www.answers.com/topic/cavalier [0.527] [2]

chips association:  www.reviewcentre.com/reviews14348.html [0.2252056] [1]
association:  www.restaurantspy.com/ uk/england/kent/no1fishandchips.htm [0.259228] [1]
association:  www.electricscotland.com/kids/stories/fishchips7.htm [0.26986] [4]
association:  www.ukwebpages.co.uk/bardsleys/web-wall.htm [0.2195352] [1]

memory association:  www.memoryx.net/memory.html [0.64348242143492] [2]
association:  Orca.co.uk [0.68272596276825] [2]
association:  www.computermemoryupgrade.net/ [0.64348242143492] [2]

net association:  www.scienceaction.asso.fr/estran/anglais/APECHE.htm [0.1713] [1]
association:  www.fishonline.org/information/methods/ [0.1713] [1]
association:  sacoast.uwc.ac.za/education/ resources/envirofacts/gillnets.htm [0.1713] [1]

paris association:  www.guidetocinema.com/paris.html [0.42763640888889] [4]
association:  www.city-data.com/city/Paris-Texas.html [0.43780686222222] [2]
association:  lone-star.net/mall/txtrails/paris.htm [0.43780686222222] [2]

pen association:  www.mcmurrayhatchery.com/ product/portable-chicken-pen.html [0.52444] [3]
association:  www.websterspenshop.co.uk/ [0.4886] [1]
association:  www.needapresent.com/shop/ get-ProductDetail.asp?PID=1155 [0.4886] [1]

port association:  www.abports.co.uk/ [0.18304133333333] [3]
association:  www.abports.co.uk/southampton/ [0.18304133333333] [3]
association:  www.doverport.co.uk/ [0.17417733333333] [3]

tick association:  www.answers.com/topic/tick [0.668296] [2]
association:  www.amazon.com/exec/obidos/ tg/detail/-/0425167054?v=glance [0.65948848] [3]
association:  www.investordictionary.com/definition/Tick.aspx [0.676792] [2]

SESSION (426e2f2e1910da13e5f51ab5b8e3e2ab)
bug association:  onlinedictionary.datasegment.com/word/bug [0.62881422717583] [2]
association:  encyclopedia.laborlawtalk.com/Computer-bug [0.59236995043983] [2]
association:  support.microsoft.com/support/kb/articles/Q259/5/24.asp [0.48431563349333] [5]

cavalier association:  www.cavweb-forums.co.uk/index.php [0.705] [1]
association:  www.carpages.co.uk/info/vauxhall.asp [0.705] [1]
association:  www.vectra-sport.com/ [0.705] [5]

chips association:  www.thefoodplace.co.uk/restaurants/ 19876/Best Fish and Chips in Waltham Cross/ [0.266316] [1]
association:  www.electricscotland.com/kids/stories/fishchips7.htm [0.26986] [4]
association:  www.reviewcentre.com/reviews14348.html [0.2252056] [1]
association:  www.restaurantspy.com/ uk/england/kent/no1fishandchips.htm [0.259228] [1]
association:  www.kelkoo.co.uk/b/a/sbs/uk/gadgets/ keywords/poker chips/type/100138013.html [0.697798] [3]
association:  www.pokerlistings.com/poker-chips [0.697798] [3]
association:  www.ukwebpages.co.uk/bardsleys/web-wall.htm [0.2195352] [1]

fly association:  www.troutmaster.eclipse.co.uk/ [0.34771498097778] [3]
association:  www.weneedyou.com/clark-bugs/fly.html [0.3491863552] [2]
association:  www.sel.barc.usda.gov/Diptera/flies.htm [0.34903921777778] [2]
association:  www.planet-pets.com/plntfly.htm [0.35584918186667] [2]
association:  www.ivyhall.district96.k12.il.us/ 4th/kkhp/1insects/fly.html [0.35736775111111] [2]
association:  www.tacklebargains.co.uk/ [0.34833999113482] [3]

net association:  www.scienceaction.asso.fr/estran/anglais/APECHE.htm [0.1713] [1]
association:  www.worldseafishing.com/ [0.1713] [4]
```

```
association:  sacoast.uwc.ac.za/education/ resources/envirofacts/gillnets.htm [0.1713] [1]

paris association:  www.imdb.com/name/nm0385296/ [0] [5]
association:  worldfacts.us/France-La-Rochelle.htm [0] [4]
association:  Website of the city of Paris [0.127552] [1]

port association:  www.goodcooking.com/portinfo.htm [0.3389] [2]
association:  www.nfl.com/players/playerpage/302215 [0] [5]
association:  www.cellartastings.com/en/wine-feature-port-wine.html [0.33661013333333] [2]

tick association:  www.ticketmaster.co.uk/ [0] [5]
association:  bizrate.lycos.com/buy/products- -at-id286011--286070-,cid--8099.html [0.652012] [3]
association:  www.answers.com/topic/uptick [0.6775] [2]
association:  www.specialinvestor.com/terms/947.html [0.676792] [2]
association:  www.answers.com/topic/tick [0.668296] [2]

SESSION (4726102de8f4a215e4d3eb6c3b2d235d)
bug association:  www.bartleby.com/65/bu/BugPol.html [0.6574] [1]
association:  www.britannica.com/ebc/article?tocId=9358240 [0.6574] [1]
association:  encarta.msn.com/Bug-(river).html [0.6574] [1]
association:  poland.pl/nature/regions/niziny-srp/bug/description.htm [0.6574] [1]

cavalier association:  dict.die.net/cavalier/ [0.527] [2]
association:  www.answers.com/topic/cavalier [0.527] [2]
association:  www.wordwebonline.com/en/CAVALIER [0.527] [2]

chips association:  www.casino-shop.co.uk/ [0.740326] [3]
association:  www.pokershopping.co.uk/poker-chips [0.7483] [3]
association:  www.gamble.co.uk/delivery.htm [0.697798] [3]
association:  www.pokerlistings.com/poker-chips [0.697798] [3]
association:  www.kelkoo.co.uk/b/a/sbs/uk/gadgets/ keywords/poker chips/type/100138013.html [0.697798] [3]

fly association:  www.troutmaster.eclipse.co.uk/ [0.34771498097778] [3]
association:  www.fly-fishing-flies.com/ [0.34778475557926] [3]
association:  www.tacklebargains.co.uk/ [0.34833999113482] [3]
association:  www.weneedyou.com/clark-bugs/fly.html [0.3491863552] [2]
association:  www.sel.barc.usda.gov/Diptera/flies.htm [0.34903921777778] [2]
association:  www.planet-pets.com/plntfly.htm [0.35584918186667] [2]
association:  www.ivyhall.district96.k12.il.us/ 4th/kkhp/1insects/fly.html [0.35736775111111] [2]
association:  www.fliesonline.co.uk/ [0.34806681524149] [3]

memory association:  www.sciencedaily.com/releases/2003/11/031112073642.htm [0.22257277333333] [3]
association:  www.medicalnewstoday.com/medicalnews.php?newsid=15950 [0.1649] [3]
association:  www.cat.cc.md.us/courses/bio141/lecguide/ unit3/humoral/antibodies/memory/memory.html [0.29714514143492] [3]
association:  www.emoryhealthcare.org/press-room/ ehc-news/2004/Nov/Long-term-Immune-Memory-Cells.html [0.28922962427937] [3]

net association:  www.internetvalley.com/intval.html [0.6505] [3]
association:  www.davesite.com/webstation/net-history.shtml [0.63680857142857] [3]
association:  www.walthowe.com/navnet/history.html [0.63680857142857] [3]
association:  www.isoc.org/internet/history/ [0.63680857142857] [3]

paris association:  www.theparisnews.com/ [0.43949493333333] [2]
association:  www.cityofparistx.com/ [0.43949493333333] [2]
association:  lone-star.net/mall/txtrails/paris.htm [0.43780686222222] [2]
association:  www.city-data.com/city/Paris-Texas.html [0.43780686222222] [2]

port association:  www.port.ac.uk/ [0.1727] [5]
association:  www.medwayports.com/ [0.1727] [3]
association:  www.doverport.co.uk/ [0.17417733333333] [3]
association:  www.abports.co.uk/southampton/ [0.18304133333333] [3]
association:  www.abports.co.uk/ [0.18304133333333] [3]
```

```
tick association:  www.answers.com/topic/uptick [0.6775] [2]
association:  www.investordictionary.com/definition/Tick.aspx [0.676792] [2]
association:  tick Definition [0.6775] [2]
association:  www.specialinvestor.com/terms/947.html [0.676792] [2]
association:  www.answers.com/topic/tick [0.668296] [2]

SESSION (5fe554031ddda8f4efd6494a54a49116)
skate single result:  no associations

SESSION (6182cfe94e8496bc22663dec2e1088db)
tick association:  www.lyme.org/ticks/removal.html [0.60798384] [1]
association:  ohioline.osu.edu/hyg-fact/2000/2073.html [0.5843272] [1]
association:  www.answers.com/topic/tick [0.668296] [2]

SESSION (66865c454d156b2208987f83fe8856c9)
bug association:  www.riob.org/ag2000/pologne-bug.htm [0.64322208] [5]
association:  www.bugjam.co.uk/ [0.63480592491378] [4]
association:  en.wikipedia.org/wiki/Computer-bug [0.45725777905635] [2]
association:  www.ccdata.com/browse/Owners-Marques-V-Volkswagen [0.64321888] [4]
association:  www.cvwoc.co.uk/pages/links.htm [0.63480592491378] [4]

cavalier association:  www.usedcarmart.co.uk/page-25612.html [0.705] [1]
association:  carsearch.yahoo.co.uk/ctl/do/ drilldown 100354123/Vauxhall,Vauxhall|Cavalier [0.705] [1]
association:  www.carpages.co.uk/info/vauxhall.asp [0.705] [1]

chips association:  www.kelkoo.co.uk/b/a/sbs/uk/gadgets/ keywords/poker chips/type/100138013.html [0.697798] [3]
association:  www.pokerlistings.com/poker-chips [0.697798] [3]
association:  www.pokershopping.co.uk/poker-chips [0.7483] [3]

memory association:  www.emoryhealthcare.org/press-room/ ehc-news/2004/Nov/Long-term-Immune-Memory-Cells.html [0.28922962427937] [3]
association:  www.liutilities.com/products/ wintaskspro/whitepapers/paper1/ [0.187448] [4]
association:  www.helpguide.org/aging/improving-memory.htm [0.17135039822222] [1]
association:  www.medicalnewstoday.com/medicalnews.php?newsid=15950 [0.1649] [3]
association:  www.memora.com/ [0.1649] [5]
association:  www.scs.tamu.edu/selfhelp/elibrary/memory.asp [0.1844416] [1]
association:  www.mindtools.com/memory.html [0.17135039822222] [1]

net association:  www.worldseafishing.com/ [0.1713] [4]
association:  www.scienceaction.asso.fr/estran/anglais/APECHE.htm [0.1713] [1]
association:  www.fishonline.org/information/methods/ [0.1713] [1]

paris association:  www.guidetocinema.com/paris.html [0.42763640888889] [4]
association:  www.city-data.com/city/Paris-Texas.html [0.43780686222222] [2]
association:  lone-star.net/mall/txtrails/paris.htm [0.43780686222222] [2]

port association:  www.abports.co.uk/ [0.18304133333333] [3]
association:  www.dotukdirectory.co.uk/ Transport/Ships-and-Shipping/Ports/ [0.1727] [3]
association:  www.medwayports.com/ [0.1727] [3]

tick association:  www.answers.com/topic/tick [0.668296] [2]
association:  www.specialinvestor.com/terms/947.html [0.676792] [2]
association:  www.investordictionary.com/definition/Tick.aspx [0.676792] [2]

SESSION (7872c7ed1fae4c24f22d2592a2ad7f50)
bug association:  en.wikipedia.org/wiki/Computer-bug [0.45725777905635] [2]
association:  www.spy.th.com/audiocat.html [0.2200548288] [3]
association:  www.ozspy.com.au/audio/audio.asp [0.1779058624] [3]
association:  uk.bizrate.com/buy/products--cat- id--14010200,keyword--Listening Devices.html [0.175271552] [3]

fly association:  www.ivyhall.district96.k12.il.us/ 4th/kkhp/1insects/fly.html [0.35736775111111] [2]
association:  www.planet-pets.com/plntfly.htm [0.35584918186667] [2]
association:  www.sewing.org/enthusiast/html/el-flyfrontzipper.html [0.51637333333333] [1]
```

```
memory association:  www.memorex.com/ [0.6504336] [5]
association:  mamory.sourceforge.net/ [0.27764] [5]
association:  www.scs.tamu.edu/selfhelp/elibrary/memory.asp [0.1844416] [1]
association:  www.emoryhealthcare.org/press-room/ ehc-news/2004/Nov/Long-term-Immune-Memory-Cells.html [0.28922962427937] [3]
association:  www.cat.cc.md.us/courses/bio141/lecguide/ unit3/humoral/antibodies/memory/memory.html [0.29714514143492] [3]
association:  www.liutilities.com/products/ wintaskspro/whitepapers/paper1/ [0.187448] [4]

net association:  www.scienceaction.asso.fr/estran/anglais/APECHE.htm [0.1713] [1]
association:  www.fishonline.org/information/methods/ [0.1713] [1]
association:  www7.taosnet.com/platinum/ data/light/whatis/fishing.html [0.1713] [1]

paris association:  www.guidetocinema.com/paris.html [0.42763640888889] [4]
association:  www.newyorksocialdiary.com/list/129.php [0.3608] [5]
association:  www.conciergerie.com/foryou/ attractions/paris/metro-map.html [0.165376] [1]
association:  www.hugahorse.org/DA/ DAP--white-plaster-of-Paris--dry-mix---4-lbs..html [0.42489066666667] [3]
association:  doityourself.com/patch/index.shtml [0.3608] [4]
association:  www.misterart.com/store/view/ 003/group-id/965/DAP-Plaster-of-Paris.htm [0.42489066666667] [3]

pen association:  www.cpen.com/ [0.4886] [5]
association:  www.merlynspen.com/ [0.52444] [5]
association:  www.suburbansurgical.com/custompen.html [0.52444] [3]
association:  www.needapresent.com/shop/ get-ProductDetail.asp?PID=1155 [0.4886] [1]
association:  www.pen.org/ [0.4886] [5]

port association:  www.abports.co.uk/ [0.18304133333333] [3]
association:  www.abports.co.uk/southampton/ [0.18304133333333] [3]
association:  www.the-port-man.fsbusiness.co.uk/ [0.32242773333333] [2]

tick association:  www.amazon.com/exec/obidos/ tg/detail/-/B0000AUHQE?v=glance [0.6078328] [4]
association:  www.anapsid.org/lyme/howtoremoveticks.pdf [0.580751328] [1]
association:  www.lyme.org/ticks/removal.html [0.60798384] [1]
association:  ohioline.osu.edu/hyg-fact/2000/2073.html [0.5843272] [1]

SESSION (949d488fb11092350b57cfd0fdcf3994)
fly association:  www.tacklebargains.co.uk/ [0.34833999113482] [3]
association:  www.fly-fishing-flies.com/ [0.34778475557926] [3]
association:  www.ryanair.com/site/EN/ [0.3472] [4]

net association:  www.felixonline.co.uk/v2/article.php?id=2268 [0.25344857142857] [4]
association:  sacoast.uwc.ac.za/education/ resources/envirofacts/gillnets.htm [0.1713] [1]
association:  www.scienceaction.asso.fr/estran/anglais/APECHE.htm [0.1713] [1]

SESSION (ade4f0fd40a69d21810c682362ce52c8)
fly association:  www.fly-fishing-flies.com/ [0.34778475557926] [3]
association:  www.tacklebargains.co.uk/ [0.34833999113482] [3]

SESSION (ba67b0105e41e4acb0ec94114829962e)
bug association:  en.wikipedia.org/wiki/Computer-bug [0.45725777905635] [2]
association:  support.microsoft.com/support/kb/articles/Q259/5/24.asp [0.48431563349333] [5]
association:  encyclopedia.laborlawtalk.com/Computer-bug [0.59236995043983] [2]

cavalier association:  www.usedcarmart.co.uk/page-25612.html [0.705] [1]
association:  carsearch.yahoo.co.uk/ctl/do/ drilldown 100354123/Vauxhall,Vauxhall|Cavalier [0.705] [1]
association:  www.carpages.co.uk/info/vauxhall.asp [0.705] [1]

chips association:  www.restaurantspy.com/ uk/england/kent/no1fishandchips.htm [0.259228] [1]
association:  www.thefoodplace.co.uk/restaurants/ 19876/Best Fish and Chips in Waltham Cross/ [0.266316] [1]
association:  www.electricscotland.com/kids/stories/fishchips7.htm [0.26986] [4]
association:  www.kelkoo.co.uk/b/a/sbs/uk/gadgets/ keywords/poker chips/type/100138013.html [0.697798] [3]
association:  www.pokerlistings.com/poker-chips [0.697798] [3]
association:  www.gamble.co.uk/delivery.htm [0.697798] [3]
```

```
fly association:  www.diynet.com/diy/na-serging/ article/0,2025,DIY-14143-2277030,00.html [0.5424] [1]
association:  www.hgtv.com/hgtv/cr-sewing-buttons-zippers/ article/0,1789,HGTV-3325-1377114,00.html [0.5424] [1]
association:  sewing.patternreview.com/ cgi-bin/review/readreview.pl?ID=623 [0.5424] [4]

net association:  www.felixonline.co.uk/v2/article.php?id=2268 [0.25344857142857] [4]
association:  today.reuters.co.uk/stocks/QuoteCompanyNewsArticle.  aspx?view=CN [0.58204285714286] [4]
association:  www.wordwebonline.com/en/NETPROFIT [0.63822335238095] [2]
association:  www.hll.com/HLL/investors/glossary.html [0.63822335238095] [2]
association:  www.investordictionary.com/definition/Net profit.aspx [0.63749314285714] [2]
association:  www.isoc.org/internet/history/ [0.63680857142857] [3]
association:  www.walthowe.com/navnet/history.html [0.63680857142857] [3]
association:  www.davesite.com/webstation/net-history.shtml [0.63680857142857] [3]

paris association:  www.city-data.com/city/Paris-Texas.html [0.43780686222222] [2]
association:  lone-star.net/mall/txtrails/paris.htm [0.43780686222222] [2]
association:  www.cityofparistx.com/ [0.43949493333333] [2]
association:  www.theparisnews.com/ [0.43949493333333] [2]

port association:  www.epicurious.com/drinking/ wine-dictionary/entry?id=7570 [0.33661013333333] [2]
association:  www.cellartastings.com/en/wine-feature-port-wine.html [0.33661013333333] [2]
association:  www.the-port-man.fsbusiness.co.uk/ [0.32242773333333] [2]
association:  winemaking.jackkeller.net/reques13.asp [0.337792] [4]

tick association:  www.amazon.com/exec/obidos/ tg/detail/-/B0000AUHQE?v=glance [0.6078328] [4]
association:  www.amazon.com/exec/obidos/ tg/detail/-/0425167054?v=glance [0.65948848] [3]
association:  www.newkadia.com/NK.php?ipg=3928 [0.63773872] [3]

SESSION (bc00069c86b1d0f02da7ce81aef1db5a)
bug association:  en.wikipedia.org/wiki/Computer-bug [0.45725777905635] [2]
association:  support.microsoft.com/support/kb/articles/Q259/5/24.asp [0.48431563349333] [5]
association:  bugclub.org/beginners/software/bugs.html [0.61030676309333] [2]

cavalier association:  www.wordwebonline.com/en/CAVALIER [0.527] [2]
association:  dogs-puppies.dogs-central.com/ cavalier-king-charles-spaniel-puppy/ [0.527] [4]
association:  dict.die.net/cavalier/ [0.527] [2]

chips association:  www.reviewcentre.com/reviews14348.html [0.2252056] [1]
association:  www.restaurantspy.com/ uk/england/kent/no1fishandchips.htm [0.259228] [1]
association:  www.thefoodplace.co.uk/restaurants/ 19876/Best Fish and Chips in Waltham Cross/ [0.266316] [1]

fly association:  www.diynet.com/diy/na-serging/ article/0,2025,DIY-14143-2277030,00.html [0.5424] [1]
association:  sewing.patternreview.com/ cgi-bin/review/readreview.pl?ID=623 [0.5424] [4]
association:  www.sewnews.com/library/sewnews/qa/aaqa0702c.htm [0.5424] [1]

memory association:  www.memorex.com/ [0.6504336] [5]
association:  www.laptopshop.co.uk/laptop-memory.htm [0.63963422943492] [2]
association:  Orca.co.uk [0.68272596276825] [2]

net association:  www.wordwebonline.com/en/NETPROFIT [0.63822335238095] [2]
association:  www.hll.com/HLL/investors/glossary.html [0.63822335238095] [2]
association:  www.investordictionary.com/definition/Net profit.aspx [0.63749314285714] [2]
association:  www.investorwords.com/3259/net-profit.html [0.64698586666667] [2]

paris association:  www.guidetocinema.com/paris.html [0.42763640888889] [4]
association:  www.city-data.com/city/Paris-Texas.html [0.43780686222222] [2]
association:  lone-star.net/mall/txtrails/paris.htm [0.43780686222222] [2]

pen association:  members.tripod.com/gamebirds/ourmuteswans.htm [0.88284] [2]
association:  www.doh.gov.ph/bird-flu.htm [0.9366] [4]
association:  www.uen.org/swan/TSJigsaw.html [0.88284] [2]

port association:  www.epicurious.com/drinking/ wine-dictionary/entry?id=7570 [0.33661013333333] [2]
association:  www.cellartastings.com/en/wine-feature-port-wine.html [0.33661013333333] [2]
```

```
association:  www.the-port-man.fsbusiness.co.uk/ [0.32242773333333] [2]

tick association:  www.anapsid.org/lyme/howtoremoveticks.pdf [0.580751328] [1]
association:  www.lyme.org/ticks/removal.html [0.60798384] [1]
association:  ohioline.osu.edu/hyg-fact/2000/2073.html [0.5843272] [1]

SESSION (d7c34dee65c406a2f8c01ef2c6774177)
bug association:  www.spy.th.com/audiocat.html [0.2200548288] [3]
association:  www.ozspy.com.au/audio/audio.asp [0.1779058624] [3]
association:  www.edirectory.co.uk/pf/pages/moreinfoa.  asp?pe=FFDIJHQ- Bug em Listening Device [0.1673686208] [3]

cavalier association:  www.wordwebonline.com/en/CAVALIER [0.527] [2]
association:  www.answers.com/topic/cavalier [0.527] [2]
association:  onlinedictionary.datasegment.com/word/cavalier [0.527] [2]

fly association:  www.ivyhall.district96.k12.il.us/ 4th/kkhp/1insects/fly.html [0.35736775111111] [2]
association:  www.planet-pets.com/plntfly.htm [0.35584918186667] [2]
association:  www.sel.barc.usda.gov/Diptera/flies.htm [0.34903921777778] [2]
association:  www.weneedyou.com/clark-bugs/fly.html [0.3491863552] [2]
association:  insected.arizona.edu/flyinfo.htm [0.35181539555556] [2]
association:  www.tacklebargains.co.uk/ [0.34833999113482] [3]
association:  www.sportfish.co.uk/ [0.35153777777778] [3]
association:  www.fliesonline.co.uk/ [0.34806681524149] [3]

memory association:  www.emoryhealthcare.org/press-room/ ehc-news/2004/Nov/Long-term-Immune-Memory-Cells.html [0.28922962427937] [3]
association:  microvet.arizona.edu/Courses/ MIC419/Tutorials/bigpicture.html [0.46726729610159] [3]
association:  www.cat.cc.md.us/courses/bio141/lecguide/ unit3/humoral/antibodies/memory/memory.html [0.29714514143492] [3]
association:  www.sciencedaily.com/releases/2003/11/031112073642.htm [0.22257277333333] [3]
association:  www.laptopshop.co.uk/laptop-memory.htm [0.63963422943492] [2]
association:  www.memoryx.net/memory.html [0.64348242143492] [2]
association:  www.computermemoryupgrade.net/ [0.64348242143492] [2]

paris association:  www.conciergerie.com/foryou/ attractions/paris/metro-map.html [0.165376] [1]
association:  Website of the city of Paris [0.127552] [1]
association:  www.francetourism.com/ [0.127552] [1]

pen association:  www.cpen.com/ [0.4886] [5]
association:  buy.ebay.co.uk/parker [0.4886] [1]
association:  www.needapresent.com/shop/ get-ProductDetail.asp?PID=1155 [0.4886] [1]
association:  uk.bizrate.com/buy/products- -cat-id--212,keyword--Pen Refills.html [0.4886] [1]

port association:  www.abports.co.uk/ [0.18304133333333] [3]
association:  www.abports.co.uk/southampton/ [0.18304133333333] [3]
association:  www.doverport.co.uk/ [0.17417733333333] [3]

tick association:  www.anapsid.org/lyme/howtoremoveticks.pdf [0.580751328] [1]
association:  ohioline.osu.edu/hyg-fact/2000/2073.html [0.5843272] [1]
association:  dogs.about.com/cs/disableddogs/qt/tick-removal.htm [0.575857632] [1]

SESSION (da6c39a0038cca0445035c64a515c19f)
bug association:  www.britannica.com/ebc/article?tocId=9358240 [0.6574] [1]
association:  www.bartleby.com/65/bu/BugPol.html [0.6574] [1]
association:  poland.pl/nature/regions/niziny-srp/bug/description.htm [0.6574] [1]

cavalier association:  www.usedcarmart.co.uk/page-25612.html [0.705] [1]
association:  carsearch.yahoo.co.uk/ctl/do/ drilldown 100354123/Vauxhall,Vauxhall|Cavalier [0.705] [1]
association:  www.carpages.co.uk/info/vauxhall.asp [0.705] [1]
association:  www.cavweb-forums.co.uk/index.php [0.705] [1]
association:  www.vectra-sport.com/ [0.705] [5]

chips association:  www.reviewcentre.com/reviews14348.html [0.2252056] [1]
association:  www.restaurantspy.com/ uk/england/kent/no1fishandchips.htm [0.259228] [1]
association:  www.harryramsdens.co.uk/about/downloads/statistics.doc [0.2195352] [1]
```

```
association:  www.ukwebpages.co.uk/bardsleys/web-wall.htm [0.2195352] [1]

fly association:  www.ivyhall.district96.k12.il.us/ 4th/kkhp/1insects/fly.html [0.35736775111111] [2]
association:  www.planet-pets.com/plntfly.htm [0.35584918186667] [2]
association:  www.sel.barc.usda.gov/Diptera/flies.htm [0.34903921777778] [2]
association:  www.weneedyou.com/clark-bugs/fly.html [0.3491863552] [2]
association:  insected.arizona.edu/flyinfo.htm [0.35181539555556] [2]

memory association:  www.scs.tamu.edu/selfhelp/elibrary/memory.asp [0.1844416] [1]
association:  www.helpguide.org/aging/improving-memory.htm [0.17135039822222] [1]
association:  www.mindtools.com/memory.html [0.17135039822222] [1]
association:  www.thememorypage.net/ [0.172416] [1]
association:  familydoctor.org/124.xml [0.172416] [4]
association:  webster.commnet.edu/faculty/ simonds/memory.htm [0.172416] [1]

net association:  www.wordwebonline.com/en/NETPROFIT [0.63822335238095] [2]
association:  www.hll.com/HLL/investors/glossary.html [0.63822335238095] [2]
association:  www.investorwords.com/3259/net-profit.html [0.64698586666667] [2]

paris association:  www.city-data.com/city/Paris-Texas.html [0.43780686222222] [2]
association:  www.cityofparistx.com/ [0.43949493333333] [2]
association:  www.theparisnews.com/ [0.43949493333333] [2]
association:  lone-star.net/mall/txtrails/paris.htm [0.43780686222222] [2]

pen association:  members.tripod.com/gamebirds/ourmuteswans.htm [0.88284] [2]
association:  www.bbc.co.uk/nature/wildfacts/factfiles/3029.shtml [0.9366] [2]
association:  www.city.stratford.on.ca/ site-tourism/about-stratford-swans.asp [0.9366] [2]
association:  www.nps.gov/yell/nature/animals/birds/trumpeter.htm [0.9366] [2]
association:  www.uen.org/swan/TSJigsaw.html [0.88284] [2]

port association:  www.detto.com/learningcenter/MLCTOTMoct2003.htm [0.4401] [1]
association:  computer.howstuffworks.com/usb.htm [0.4401] [1]
association:  www.computerhope.com/network/ports.htm [0.4401] [1]

tick association:  www.amazon.com/exec/obidos/ tg/detail/-/0425167054?v=glance [0.65948848] [3]
association:  www.newkadia.com/NK.php?ipg=3928 [0.63773872] [3]
association:  www.mycomicshop.com [0.652012] [3]
association:  bizrate.lycos.com/buy/products-at-id286011--286070-,cid--8099.html [0.652012] [3]

SESSION (df4e4c71a31a1613b3fa3562d616cd1c)
bug association:  en.wikipedia.org/wiki/Computer-bug [0.45725777905635] [2]
association:  encyclopedia.laborlawtalk.com/Computer-bug [0.59236995043983] [2]
association:  onlinedictionary.datasegment.com/word/bug [0.62881422717583] [2]

chips association:  www.oldsoftware.com/z80.html [0.006] [2]
association:  www.tigerdirect.com/ [0.006] [2]
association:  www.windowsitlibrary.com/Content/175/03/2.html [0.006] [4]

fly association:  www.tacklebargains.co.uk/ [0.34833999113482] [3]
association:  www.planet-pets.com/plntfly.htm [0.35584918186667] [2]
association:  www.weneedyou.com/clarkbugs/fly.html [0.3491863552] [2]
association:  www.iflyamerica.com/ [0.34977490488889] [4]
association:  www.fly-fishing-flies.com/ [0.34778475557926] [3]
association:  www.troutmaster.eclipse.co.uk/ [0.34771498097778] [3]

memory association:  www.emoryhealthcare.org/pressroom/ ehcnews/2004/Nov/Long-term-Immune-Memory-Cells.html [0.28922962427937] [3]
association:  www.helpguide.org/aging/improving-memory.htm [0.17135039822222] [1]
association:  www.mindtools.com/memory.html [0.17135039822222] [1]

paris association:  www.hugahorse.org/DA/ DAP--white-plaster-of-Paris--dry-mix---4-lbs..html [0.42489066666667] [3]
association:  www.misterart.com/store/view/ 003/group-id/965/DAP-Plaster-of-Paris.htm [0.42489066666667] [3]
association:  familycrafts.about.com/ cs/miscjewelry/a/blplastercast.htm [0.43749866666667] [3]
```

```
pen association:  www.mcmurrayhatchery.com/ product/portable-chicken-pen.html [0.52444] [3]
association:  www.tortoise.org/general/pen.html [0.6678] [3]
association:  www.suburbansurgical.com/custompen.html [0.52444] [3]
association:  members.tripod.com/gamebirds/ourmuteswans.htm [0.88284] [2]
association:  www.merlynspen.com/ [0.52444] [5]
association:  www.uen.org/swan/TSJigsaw.html [0.88284] [2]

port association:  computer.howstuffworks.com/usb.htm [0.4401] [1]
association:  www.computerhope.com/network/ports.htm [0.4401] [1]
association:  www.abilityhub.com/information/ports.htm [0.4401] [1]

tick association:  www.amazon.com/exec/obidos/ tg/detail/-/B0000AUHQE?v=glance [0.6078328] [4]
association:  www.amazon.com/exec/obidos/ tg/detail/-/0425167054?v=glance [0.65948848] [3]
association:  www.newkadia.com/NK.php?ipg=3928 [0.63773872] [3]

SESSION (e7a9048ec40375797e6b72e66d71fd6a)
bug association:  www.bugjam.co.uk/ [0.63480592491378] [4]
association:  www.cvwoc.co.uk/pages/links.htm [0.63480592491378] [4]
association:  www.stevenagevwclub.co.uk/links.html [0.65887082091378] [4]
association:  en.wikipedia.org/wiki/Computer-bug [0.45725777905635] [2]
association:  encyclopedia.laborlawtalk.com/Computer-bug [0.59236995043983] [2]
association:  onlinedictionary.datasegment.com/word/bug [0.62881422717583] [2]

cavalier association:  www.usedcarmart.co.uk/page-25612.html [0.705] [1]
association:  carsearch.yahoo.co.uk/ctl/do/ drilldown 100354123/Vauxhall,Vauxhall|Cavalier [0.705] [1]
association:  www.carpages.co.uk/info/vauxhall.asp [0.705] [1]
association:  www.cavweb-forums.co.uk/index.php [0.705] [1]

chips association:  www.kelkoo.co.uk/b/a/sbs/uk/gadgets/ keywords/poker chips/type/100138013.html [0.697798] [3]
association:  www.pokerlistings.com/poker-chips [0.697798] [3]
association:  www.gamble.co.uk/delivery.htm [0.697798] [3]
association:  www.casino-shop.co.uk/ [0.740326] [3]

fly association:  www.tacklebargains.co.uk/ [0.34833999113482] [3]
association:  www.fly-fishing-flies.com/ [0.34778475557926] [3]
association:  www.fliesonline.co.uk/ [0.34806681524149] [3]

memory association:  www.emoryhealthcare.org/press-room/ ehc-news/2004/Nov/Long-term-Immune-Memory-Cells.html [0.28922962427937] [3]
association:  microvet.arizona.edu/Courses/ MIC419/Tutorials/bigpicture.html [0.46726729610159] [3]
association:  www.cat.cc.md.us/courses/bio141/lecguide/ unit3/humoral/antibodies/memory/memory.html [0.29714514143492] [3]
association:  www.tufts.edu/sackler/immunology/ward/ [0] [4]
association:  www.laptopshop.co.uk/laptop-memory.htm [0.63963422943492] [2]
association:  Orca.co.uk [0.68272596276825] [2]
association:  www.memoryx.net/memory.html [0.64348242143492] [2]
association:  www.computermemoryupgrade.net/ [0.64348242143492] [2]

paris association:  www.guidetocinema.com/paris.html [0.42763640888889] [4]
association:  www.city-data.com/city/Paris-Texas.html [0.43780686222222] [2]
association:  lone-star.net/mall/txtrails/paris.htm [0.43780686222222] [2]

port association:  computer.howstuffworks.com/usb.htm [0.4401] [1]
association:  www.computerhope.com/network/ports.htm [0.4401] [1]
association:  www.abilityhub.com/information/ports.htm [0.4401] [1]

tick association:  www.anapsid.org/lyme/howtoremoveticks.pdf [0.580751328] [1]
association:  www.lyme.org/ticks/removal.html [0.60798384] [1]
association:  dogs.about.com/cs/disableddogs/qt/tick-removal.htm [0.575857632] [1]

SESSION (fd16572425d3161b0910e7b7b4cc419f)
fish single result:  no associations
```

Processing complete [June 23rd 2005 04:30:33]:

Appendix D

Publications

Duncan Martin, Mark Truran, Helen Ashman *The end-point is not enough*, The Fifteenth ACM Conference on Hypertext and Hypermedia, August 2004.

Mark Truran, Duncan Martin, Helen Ashman, *Goate, anyone?*, The Fourteenth ACM conference on Hypertext and Hypermedia, August 2003.

Duncan Martin, Mark Truran, Helen Ashman, *Caching approaches for content Altering HTTP proxies*, IEEE ICITA conference, November 2002.

Duncan Martin, Mark Truran, Helen Ashman, *Implementing Conceptual Linking on Todays Web*, Ausweb International Conference, July 2002.

Pending Decision

Mark Truran, Helen Ashman, *Automatic Hypertext Generation Systems*, submitted to ACM Computing Surveys June 29th 2004. Manuscript ID CSUR-2004-0013.

Mark Truran, Helen Ashman, *Relevance Feedback in Information Retrieval*, submitted to ACM Computing Surveys September 9th, 2004. Manuscript ID CSUR-2004-0059. Paper revised and re-submitted as *Key Aspects of the Relevance Feedback Process* CSUR-2004-0059.R1 on April 4, 2005.

Mark Truran, James Goulding, Helen Ashman, *Co-Active Intelligence for Image Retrieval*, submitted to ACM Multimedia 2005 on 30th May 2005.

References

[1] Html 4.01 specification. http://www.w3.org/TR/REC-html40/, December 1999. W3C.

[2] flickr beta. http://www.flickr.com/, 2005. Yahoo! Inc.

[3] M. Agosti and J. Allan. Methods and tools for the construction of hypertexts. *Information Processing and Management*, 33(2):129–271, 1997.

[4] M. Agosti and A. Smeaton, editors. *Information Retrieval and Hypertext*. Kluwer Academic Publishers, Boston, USA, 1996.

[5] A. Aizawa. An information-theoretic perspective of tf-idf measures. *Information Processing and Management*, 39(1):45–65, 2003.

[6] J. Allan. *Automatic Hypertext Construction*. PhD thesis, Cornell University, USA, January 1995.

[7] J. Allan. Automatic hypertext link typing. In *Proceedings of the the seventh ACM conference on Hypertext*, pages 42–52, New York, NY, USA, 1996. ACM Press.

[8] J. Allan. Hard track overview in trec 2003. In E. M. Voorhees and L. P. Buckland, editors, *The Twelfth Text Retrieval Conference(TREC 2003)*, pages 24–38, Gaithersburg, Maryland, USA, 2003. National Institute of Standards and Technology. NIST Special Publication 500-255.

[9] R. B. Allen, P. Obry, and M. Littman. An interface for navigating clustered document sets returned by queries. In *COCS '93: Proceedings of the conference*

on Organizational computing systems, pages 166–171, New York, NY, USA, 1993. ACM Press.

[10] M. R. Anderberg. *Cluster Analysis for Applications*. Academic Press Inc., New York, NY, USA, 1973.

[11] M. H. Anderson, J. Nielsen, and H. Rasmussen. A similarity-based hypertext browser for reading the unix network news. *Hypermedia*, 1(3):255–265, 1989.

[12] V. Apparao, S. Byrne, M. Champion, S. Isaacs, I. Jacobs, A. Le Hors, G. Nicol, J. Robie, R. Sutor, C. Wilson, and L. Wood. Document object model (DOM) level 1 specification. http://www.w3.org/TR/REC-DOM-Level-1/, October 1998. W3C.

[13] T. Asano, B. Bhattacharya, M. Keil, and F. Yao. Clustering algorithms based on minimum and maximum spanning trees. In *Proceedings of the fourth annual symposium on Computational geometry*, pages 252–257, New York, NY, USA, 1988. ACM Press.

[14] H. Ashman. *Theory and Practice of Large-Scale Hypermedia Management Systems*. PhD thesis, The Royal Melbourne Institute of Technology University, Australia, 1999.

[15] H. Ashman. Electronic document addressing: dealing with change. *ACM Computing Surveys*, 32(3):201–212, 2000.

[16] H. Ashman, A. Garrido, and H. Oinas-Kukkonen. Hand-made and computed links, precomputed and dynamic links. In N. Fuhr, G. Dittrich, and K. Tochtermann, editors, *HIM 97': Proceedings of Hypertext - Information Retrieval - Multimedia Conference 1997*. Universitätsverlag Konstanz, Germany, 1997.

[17] R. Attar and A. S. Fraenkel. Local feedback in full-text retrieval systems. *Journal of the ACM*, 24(3):397–417, 1977.

[18] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co. Inc., Boston, MA, 1999.

[19] P. Bailey, N. Craswell, and D. Hawking. Dark matter on the web. In *9th International World Wide Web Conference (Poster Session)*. Foretec Seminars, 2000. ISBN 1-930792-00-X.

[20] J. Ballerini, M. Buchel, D. Knaus, B. Mateev, P. Mittendorf, P. Schauble, P. Sheridan, and M. Wechsler. Spider retrieval system at trec-5. In *The Fifth text Retrieval Conference (TREC-5)*, pages 217–229, Gaithersburg, MA, USA, 1996. National Institute of Science and Technology. NIST Special Publication 500-238.

[21] M. Beaulieu. Experiments with interfaces to support query expansion. *Journal of Documentation*, 53(1):8–19, 1997.

[22] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 407–416, New York, NY, USA, 2000. ACM Press.

[23] N. J. Belkin, I. Chaleva, M. Cole, Y.-L. Li, L. Liu, Y.-H. Liu, G. Muresan, C. L. Smith, Y. Sun, X.-J. Yuan, and X.-M. Zhang. Rutgers' hard track experiences at trec 2004. In E. M. Voorhees and L. P. Buckland, editors, *The Thirteenth Text Retrieval Conference(TREC 2004)*, Gaithersburg, Maryland, USA, 2004. National Institute of Standards and Technology. NIST Special Publication 500-261.

[24] S. Benford, D. Snowdon, C. Greenhalgh, R. Ingram, and I. Know. VR-VIBE: A virtual environment for co-operative information retrieval. *Computer Graphics Forum (Proceedings Eurographics'95)*, 14(2):349–360, 1995.

[25] M. Bernstein. An apprentice that discovers hypertext links. In A. Rizk, N. Streitz, and J. André, editors, *Hypertext: concepts, systems and applications*, pages 212–223. Cambridge University Press, New York, NY, USA, 1992.

[26] P. V. Biron and D. H. Kraft. New methods for relevance feedback: Improving

information retrieval performance. In *Proceedings of the 1995 ACM symposium on Applied computing*, pages 482–487, New York, NY, USA, 1995. ACM Press.

[27] J. Blustein. Automatically generated hypertext versions of scholarly articles and their evaluation. In *Proceedings of the eleventh ACM on Hypertext and Hypermedia*, pages 201–210, New York, NY, USA, 2000. ACM Press.

[28] A. Bookstein. Fussy requests: An approach to weighted boolean searches. *Journal of the American Society for Information Science*, 31(4):275–279, 1980.

[29] A. Bookstein. Information retrieval: A sequential learning process. *Journal of the American Society for Information Science*, 34(5):331–342, 1983.

[30] C. L. Borgman. Why are online catalogs hard to use? *Journal of the American Society for Information Science*, 6:387–400, 1986.

[31] C. L. Borgman. Why are online catalogs still hard to use? *J. Am. Soc. Inf. Sci.*, 47(7):493–503, 1996.

[32] D. Box, D. Ehnebuske, G. Kakivaya, A. Layman, N. Mendelsohn, H. F. Nielsen, S. Thatte, and D. Winer. Simple Object Access Protocol (SOAP) 1.1, May 2000. http://www.w3.org/TR/2000/NOTE-SOAP-20000508/.

[33] T. J. Brailsford, C. D. Stewart, M. R. Zakaria, and A. Moore. Auto-navigation, links and narrative in an adaptive web-based integrated learning environment. In *Proceedings of the Eleventh International World Wide Web Conference*, New York, NY, USA, 2002. ACM.

[34] M. Braschler, B. Mateev, E. Mittendorf, P. Schäuble, and M. Wechsler. SPIDER retrieval system at trec7. In *The Seventh Text Retrieval Conference (TREC-7)*, pages 446–454, Gaithersburg, Maryland, USA, 1998. National Institute for Science and Technology. NIST Special Publication 500-242.

[35] T. Bray, J. Paoli, C. M. Sperbourg-McQueen, E. Maler, and F. Yergeau. Extensible mark up language (xml) 1.0 (third edition). http://www.w3.org/TR/REC-xml/, February 2004. W3C.

[36] B. E. Brewington and G. Cybenko. Keeping up with the changing web. *Computer*, 33(5):52–58, 2000.

[37] S. Brin and L. Page. The Anatomy of a Large-scale Hypertextual Web Search Engine. In *WWW7: Proceedings of the seventh international conference on World Wide Web 7*, pages 107–117. Elsevier Science Publishers B. V., 1998.

[38] P. J. Brown. Interactive documentation. *Software - Practice and Experince*, 16(3):292–299, 1986.

[39] P. Browne and A. F. Smeaton. Video Information Retrieval using Objects and Ostensive Relevance Feedback. In *Proceedings of the 2004 ACM symposium on Applied computing*, pages 1084–1090, New York, NY, USA, 2004. ACM Press.

[40] R. Bruce and J. Wiebe. Word-sense Disambiguation using Decomposable Models. In *Proceedings of the 32nd conference on Association for Computational Linguistics*, pages 139–146. Association for Computational Linguistics, 1994.

[41] M. Buckland and F. Gey. The relationship between recall and precision. *Journal of the American Society for Information Science*, 45(1):12–19, 1994.

[42] C. Buckley. The importance of proper term weighting. In M. Bates, editor, *Human Language Technology*. Morgan Kaufman, San Francisco, CA, USA, 1993.

[43] C. Buckley, J. Allan, and G. Salton. Automatic routing and ad-hoc retrieval using SMART: TREC 2. In *Proceedings of the Second Text Retrieval Conference*, pages 45–56, Gaithersburg, MA, USA, 1994. National Institute for Science Technology. Special Publication 500-215.

[44] C. Buckley, M. Mitra, J. A. Walz, and C. Cardie. Using clustering and superconcepts within SMART: TREC 6. *Information Processing and Management*, 36(1):109–131, 2000.

[45] C. Buckley and G. Salton. Optimization of relevance feedback weights. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Re-*

search and Development in Information Retrieval, pages 351–357, New York, NY, USA, 1995. ACM Press.

[46] C. Buckley, G. Salton, and J. Allan. Automatic retrieval with locality information using SMART. In *The First Text REtrieval Conference (TREC-1)*, pages 59–72, Gaithersburg, MA, USA, 1992. National Institute for Science Technology. Special Publication 500-207.

[47] C. Buckley, G. Salton, and J. Allan. The effect of adding relevance information in a relevance feedback environment. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 292–300, New York, NY, USA, 1994. ACM.

[48] C. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic query expansion using SMART: TREC 3. In *Proceedings of Third Text Retrieval conference (TREC- 3)*, pages 69–81, Gaithersburg, MA, USA, 1995. National Institute for Science Technology. NIST Special Publication 500-225.

[49] C. Buckley, A. Singhal, M. Mitra, and G. Salton. New retrieval approaches using smart: Trec 4. In D. K. Harman, editor, *TREC 4 - The Fourth Text Retrieval Conference*, pages 25–49, Gaithersburg, MA, USA, November 1995. National Institute of Science and Technology. NIST Special Publication 500-236.

[50] V. Bush. As we may think. *The Atlantic Monthly*, 176(1):101–108, July 1945.

[51] V. Bush. Memex re-visited. pages 179–191. ASLIB, London, 1987.

[52] D. Cai, X. He, Z. Li, W.-Y. Ma, and J.-R. Wen. Hierarchical Clustering of WWW Image Search Results using Visual, Textual and Link Information. In *MULTIMEDIA '04: Proceedings of the 12th Annual ACM International Conference on Multimedia*, pages 952–959, New York, NY, USA, 2004. ACM Press.

[53] J. P. Callan and W. B. Croft. An approach to incorporating cbr concepts in ir systems. In *AAAI Spring Symposium Series: Case based reasoning and infor-*

mation retrieval - exploring the opportunities for information sharing, Menlo Park, CA, USA, 1993. AAAI Press.

[54] J. P. Callan, W. B. Croft, and S. M. Hardings. The INQUERY retreival system. In *DEXA 3: Proceedings of the Third International Conference on Database and Expert Systems*, pages 83–87, Wien, Germany, 1992. Springer-Verlag. ISBN 3-211-82400-6.

[55] I. Campbell and K. van Rijsbergen. The ostensive model of developing information needs. In P. Ingwersen and N. O. Pors, editors, *CoLIS 2: Second International Conference on Conceptions of Library and Information Science*, Copenhagen, Denmark, October 1996. School of Librarianship.

[56] F. Can. Incremental clustering for dynamic information processing. *ACM Transactions on Information Processing*, 11(2):143–164, 1993.

[57] C. Carpineto, R. de Mori, G. Romano, and B. Bigi. An information-theoretic approach to automatic query expansion. *ACM Trans. Inf. Syst.*, 19(1):1–27, 2001.

[58] C. Carpineto, G. Romano, and V. Giannini. Improving retrieval feedback with multiple term-ranking function combination. *ACM Trans. Inf. Syst.*, 20(3):259–290, 2002.

[59] J. W. T. Cawley. *Flinkman*. Honours Thesis, Discipline of Computer Science,, 1993. Flinders Univeristy of South Australia.

[60] S.-F. Chang, J. R. Smith, M. Beigi, and A. Benitez. Visual Information Retrieval from Large Distributed Online Repositories. *Communications of the ACM*, 40(12):63–71, 1997.

[61] Y. Chen, J. Z. Wang, and R. Krovetz. Content-based image retrieval by clustering. In *MIR '03: Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, pages 193–200, New York, NY, USA, 2003. ACM Press.

[62] C. H. Cheng. A branch and bound clustering algorithm. *IEEE Transactions on Systems, Man, and Cybernetics*, 25:895–898, 1995.

[63] J. Cho and H. Garcia-Molina. The evolution of the web and implications for an incremental crawler. In *VLDB '00: Proceedings of the 26th International Conference on Very Large Data Bases*, pages 200–209, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[64] J. Cho and H. Garcia-Molina. Estimating frequency of change. *ACM Trans. Inter. Tech.*, 3(3):256–290, 2003.

[65] C. W. Cleverdon. Report on the testing and analysis of an investigation into the comparitve efficiency of indexing systems. Technical report, College of Aeronautics, Cranfield, England, 1962.

[66] C. W. Cleverdon. The Cranfield tests on indexing language devices. *ASLIB Proceedings*, 19(6):173–194, 1967.

[67] C. W. Cleverdon. On the inverse relationship of recall and precision. *Journal of Documentation*, 28:195–201, 1972.

[68] C. W. Cleverdon, J. Mills, and E. M. Keen. Factors determining the performance of indexing systems. Technical report, Aslib Cranfield Research project, Cranfield, England, 1966.

[69] P. Cobley and L. Jansz. *Introducing Semiotics*. Icon Books, 2004.

[70] J. Conklin. Hypertext: An introduction and survey. *IEEE Computing Society Press*, 20(9):17–41, 1987.

[71] W. S. Cooper. Some inconsistencies and misnomers in probabilistic information retrieval. In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 57–61. ACM Press, 1991.

[72] W. S. Cooper and P. Huizinga. The maximum entropy principle and its application to the design of probabilistic retrieval systems. *Information technology: Research and Development*, 1:99–112, 1982.

[73] F. Crestani, M. Lalmas, C. J. Van Rijsbergen, and I. Campbell. Is this document relevant? ...probably: A survey of probabilistic models in information retrieval. *ACM Comput. Surv.*, 30(4):528–552, 1998.

[74] W. B. Croft. Clustering large files of documents using the single-link method. *Journal of the American Society for Information Science*, 28:341–344, 1977.

[75] W. B. Croft. *Organizing and Searching Large Files of Document Descriptions*. PhD thesis, Churchill College, University of Cambridge, 1978.

[76] W. B. Croft. Incorporating different search models into one document retrieval system. In *Proceedings of the 4th annual international ACM SIGIR conference on Information storage and retrieval*, pages 40–45, New York, NY, USA, 1981. ACM Press.

[77] W. B. Croft, R. Cook, and D. Wilder. Providing Government Information on the Internet: Experiences with THOMAS. In *Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries (DL '95)*, pages 19–24, Austin, TX, USA, 1995. Texas A&M University.

[78] W. B. Croft and D. J. Harper. Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35:285–295, 1979.

[79] D. B. Crouch, C. J. Crouch, and G. Andreas. The Use of Cluster Hierarchies in Hypertext Information Retrieval. In *HYPERTEXT '89: Proceedings of the Second Annual ACM Conference on Hypertext*, pages 225–237. ACM Press, 1989.

[80] J. Cugini and S. Laskowski. Document clustering in concept space: The nist information retrieval visualization engine (nirve). In *ATA Euro-American Workshop on Visualization of Information and Data*. National Institute of Stan-

dards and Technology (NIST), 1997. http://zing.ncsl.nist.gov/ cugini/uicd/cc-paper.html.

[81] I. Dagan, S. Marcus, and S. Markovitch. Contextual Word Similarity and Estimation from Sparse Data. In *Proceedings of the 31st Conference on Association for Computational Linguistics*, pages 164–171. Association for Computational Linguistics, 1993.

[82] C. Davidson. What your database hides away. *New Scientist*, pages 28–31, 1993. January 9th.

[83] H. Davis, W. Hall, I. Heath, G. Hill, and R. Wilkins. Towards an integrated information environment with open hypermedia systems. In *ECHT'92: European Conference on Hypertext Technology*, pages 181–190, New York, NY, USA, December 1992. ACM Press.

[84] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

[85] A. Dempster, N. Laird, and D. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

[86] S. DeRose, R. Daniel, P. Grosso, E. Maler, J. Marsh, and N. Walsh. Xml pointer language (xpointer). http://www.w3.org/TR/xptr/, August 2002. W3C.

[87] S. DeRose, E. Maler, and D. Orchard. Xml linking language (xlink) version 1.0. http://www.w3.org/TR/xlink/, June 2001. W3C.

[88] S. J. DeRose. Xml linking. *ACM Computing Surveys (CSUR)*, 31(4es), December 1999. Article No. 21.

[89] T. E. Doszkocs. Aid, an associative interative dictionary for online spelling. *Online Review*, 2(2):163–172, 1978.

[90] T. E. Doszkocs, J. Reggia, and X. Lin. Connectionist models and information retrieval. In M. E. Williams, editor, *Annual Review of Information Science and Technology*, volume 25, pages 209–260, Amsterdam, 1990. Elsevier Science Publishers B.V.

[91] R. C. Dubes. How many clusters are best-an experiment. *Pattern Recogn.*, 20(6):645–663, 1987.

[92] B. S. Duran and P. L. Odell. *Cluster Analysis: A Survey*. Springer-Verlag, New York, NY, USA, 1974.

[93] E. N. Efthimiadis. *Interactive Query Expansion and Relevance Feedback for Document Retrieval Systems*. PhD thesis, City University, London, U.K, 1992.

[94] E. N. Efthimiadis. A user-centred evaluation of ranking algorithms for interactive query expansion. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 146–159, New York, NY, USA, 1993. ACM Press.

[95] D. Ellis, J. Furner-Hines, and P. Willett. On the creation of hypertext links in full-text documents: Measurement of inter-linker consistency. *Journal of Documentation*, 50:67–98, 1994.

[96] D. Ellis, J. Furner-Hines, and P. Willett. On the creation of hypertext links in full-text documents: measurement of retrieval effectiveness. *Journal of the American Society for Information Science*, 47(4):287–300, 1996.

[97] D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A large-scale study of the evolution of web pages. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 669–678, New York, NY, USA, 2003. ACM Press.

[98] L. Fitzpatrick and M. Dent. Automatic Feedback using Past Queries: Social Searching? In *Proceedings of the 20th Annual International ACM SIGIR Con-*

ference on Research and Development in Information Retrieval, pages 306–313, New York, NY, USA, 1997. ACM Press.

[99] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by Image and Video Content: The QBIC System. *Computer*, 28(9):23–32, 1995.

[100] C. L. Foss. Effective browsing in hypertext. In *RIAO'88: Proceedings of the Conference on User-Oriented Context-based Text and Image Handling*, pages 82–98, Paris, France, 1988. Le centre de Hautes Etudes Internationales d'Informatique Documentaire.

[101] A. M. Fountain, W. Hall, I. Heath, and H. Davis. MICROCOSM: An open model for hypermedia with dynamic linking. In A. Rizk, N. A. Streitz, and J. André, editors, *ECHT'91: Hypertext: Concepts, systems and applications: Proceedings of European Conference on Hypertext*, pages 298–311, Cambrisge, UK, 1991. Cambridge University Press.

[102] C. Fox. Lexical analysis and stoplists. In R. Baeza-Yates and W. B. Frakes, editors, *Information Retrieval: Data Structures and Algorithms*. Prentice Hall, Englewood Cliffs, 1992.

[103] C. Fox. A stop list for general text. *SIGIR Forum*, 24(1-2):19–21, r 90.

[104] L. Francisco-Revilla, F. Shipman, R. Furuta, U. Karadkar, and A. Arora. Managing change on the web. In *JCDL '01: Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, pages 67–76, New York, NY, USA, 2001. ACM Press.

[105] I. Frank M. Shipman and C. C. Marshall. Formality Considered Harmful: Experiences, Emerging Themes, and Directions on the Use of Formal Representations in Interactive Systems. *Computer Supported Cooperative Work (CSCW)*, 8(4):333–352, 1999.

[106] C. H. Franke, III and N. J. Wahl. Authoring a hypertext unix help manual. In *Proceedings of the 1995 ACM twenty-third annual conference on Computer science*, pages 238–245, New York, NY, USA, 1995. ACM Press.

[107] M. E. Frisse. Searching for information in a hypertext medical handbook. *Communications of the ACM*, 31(7):880–886, 1988.

[108] N. Fuhr. Probabilistic models in information retrieval. *The Computer Journal*, 35(3):243–255, 1992.

[109] F. Fukumoto and Y. Suzuki. Word Sense Disambiguation in Untagged Text based on Term Weight Learning. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 209–216. Association for Computational Linguistics, 1999.

[110] M. Fuller, E. Mackie, R. Sacks-Davis, and R. Wilkinson. Structured answers for a large structured document collection. In *Proceedings of the sixteenth annual international ACM SIGIR conference on Research and development in information retrieval*, pages 204–213, New York, NY, USA, 1993. ACM Press.

[111] B. V. Funt and G. D. Finlayson. Color Constant Color Indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5):522–529, 1995.

[112] R. Furuta. Concepts and models for structured documents. In J. Andre, R. Furuta, and V. Quint, editors, *Structured Documents*, pages 7–38. Cambridge University Press, Cambridge, UK, 1989.

[113] R. Furuta, C. Plaisant, and B. Shneiderman. Automatically transforming regularly structured documents into hypertext. *Electronic Publishing - Origination,Dissemination, and Design*, 2(4):211–229, 1989.

[114] R. Furuta, C. Plaisant, and B. Shneiderman. A spectrum of automatic hypertext constructions. *Hypermedia*, 1(2):179–195, 1989.

[115] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees. The TREC-8 spoken document retrieval track: A success story. In *TREC-8: Proceedings of the*

Eigth Text Retrieval Conference, pages 107–129, Gaithersburg, MA, USA, 2000. National Institute of Science and Technology. NIST Special Publication 500-246.

[116] S. Gauch and J. B. Smith. Search improvement via automatic query reformulation. *ACM Trans. Inf. Syst.*, 9(3):249–280, 1991.

[117] S. Geman and G. Geman. Stochastic Relaxation, Gibbs Distribuition and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[118] R. J. Glushko. Design issues for multi-document hypertexts. In *Proceedings of the Second Annual ACM Conference on Hypertext*, pages 51–60, New York, NY, USA, 1989. ACM Press.

[119] Google Web APIs Reference, 2004. http://www.google.com/apis/reference.html.

[120] S. J. Green. Building hypertext links by computing semantic similarity. *IEEE Transactions on Knowledge and Data Engineering*, 11(5):713–730, 1999.

[121] S. L. Greene, S. J. Devlin, P. E. Cannata, and L. M. Gomez. No ifs, ands or ors: A study of database querying. *International Journal of Man-Machine Studies*, 32(3):303–326, 1990.

[122] J. A. Guthrie, L. Guthrie, Y. Wilks, and H. Aidinejad. Subject-dependent Co-occurrence and Word Sense Disambiguation. In *Proceedings of the 29th conference on Association for Computational Linguistics*, pages 146–152. Association for Computational Linguistics, 1991.

[123] W. Hall, G. Hill, and H. Davis. The microcosm link service. In *Proceedings of the fifth ACM conference on Hypertext*, pages 256–259, New York, NY, USA, 1993. ACM Press.

[124] M. H. Hansen and E. Shriver. Using Navigation Data to Improve IR Functions in the Context of Web Search. In *CIKM '01: Proceedings of the Tenth Interna-*

tional Conference on Information and Knowledge Management, pages 135–142. ACM Press, 2001.

[125] D. Harman. *Relevance feedback and other query modification techniques*, pages 241–263. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1992.

[126] D. Harman. Relevance Feedback Revisited. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 1–10, New York, NY, USA, 1992. ACM Press.

[127] D. Harman. Overview of the second text retrieval conference (trec-2). In *Proceedings of the second conference on Text retrieval conference*, pages 271–289, New York, NY, USA, 1995. Pergamon Press, Inc.

[128] D. K. Harman. Towards interactive query expansion. In Y. Chinramella, editor, *SIGIR '88: Proceedings of the eleventh International Conference on Research and Development in Information Retrieval*, pages 321–332, New York, NY, USA, 1988. ACM.

[129] D. K. Harman. How effective is suffixing? *Journal of the American Society for Information Science*, 42(1):7–51, 1991.

[130] D. K. Harman. The DARPA tipster project. *SIGIR Forum*, 26(2):26–28, 1992.

[131] D. K. Harman, editor. *The Second Text Retrieval Conference TREC-2*, Special Publication 500-215, Gaithersburg, Maryland, USA, August 1993. National Institute of Science and Technology.

[132] D. K. Harman and C. Buckley. The nrrc reliable information access (ria) workshop. In *SIGIR '04: Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 528–529, New York, NY, USA, 2004. ACM Press.

[133] D. J. Harper. *Relevance Feedback in Document Retrieval Systems: An Evaluation of Probabilistic Strategies*. PhD thesis, Jesus College, Cambridge, UK, 1980.

[134] D. J. Harper and C. J. van Rijsbergen. An evaluation of feedback in document retrieval using co-occurence data. *Journal of Documentation*, 34:189–216, 1978.

[135] S. F. Harter. The Cranfield II relevance assessments: A critical evaluation. *Library Quarterly*, 41:229–243, 1971.

[136] J. A. Hartigan. *Clustering Algorithms*. John Wiley and Sons Inc., New York, NY, USA, 1975.

[137] D. Hawking, P. B. Thistlewaite, and N. Craswell. ANU/ACSys TREC-6 experiments. In E. Voorhees and D. K. Harman, editors, *The sixth text retrieval conference (TREC-6)*, pages 275–290, Gaithersburg, MA, USA, 1997. National Institute for Science and Technology.

[138] P. Hayes and J. Pepper. Towards an Integrated Maintenance Advisor. In *Proceedings of the Second Annual ACM Conference on Hypertext*, pages 119–127, New York, NY, USA, 1989. ACM Press.

[139] P. J. Hayes, P. Andersen, and S. Safier. Semantic case frame parsing and syntactic generality. In *Proceedings of the twenty-third Annual meeting of the Association for Computing linguistics*, Cambridge, MA, USA, June 1985. MIT Press.

[140] P. J. Hayes, L. E. Knecht, and M. J. Cellio. A News Story Categorization System. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 9–17, New York, NY, USA, 1988. ACM Press.

[141] B. He and I. Ounis. University of Glasgow at the robust track- a query-based model selection approach for the poorly-performing queries. In E. M. Voorhees and L. P. Buckland, editors, *The Twelfth Text Retrieval Conference(TREC 2003)*, pages 636–646, Gaithersburg, Maryland, USA, 2003. National Institute of Standards and Technology. NIST Special Publication 500-255.

[142] D. He and D. Demner-Fushman. Hard experiment at Maryland: From need negotiation to automated hard process. In E. M. Voorhees and L. P. Buckland,

editors, *The Twelfth Text Retrieval Conference(TREC 2003)*, pages 707–715, Gaithersburg, Maryland, USA, 2003. National Institute of Standards and Technology. NIST Special Publication 500-255.

[143] M. Hearst. Mini-paragraph segmentation of expository discourse. In *Proceedings of the 32nd Meeting of the ACL*, pages 9–16, Morristown, NJ, USA, 1994. Association for Computational Linguistics.

[144] M. A. Hearst. Texttiling: A quantitative approach to discourse. Technical report, University of California, Berkeley, USA, 1993.

[145] M. A. Hearst and C. Karadi. Cat-a-cone: an interactive interface for specifying searches and viewing retrieval results using a large category hierarchy. In *SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 246–255, New York, NY, USA, 1997. ACM Press.

[146] M. A. Hearst and J. O. Pedersen. Re-examining the Cluster Hypothesis: Scatter/gather on Retrieval Results. In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 76–84. ACM Press, 1996.

[147] G. Hirst. *Semantic Interpretation and the Resolution of Ambiguity*. Cambridge University Press, 1987.

[148] E. Ide. New experiments in relevance feedback. In G. Salton, editor, *The SMART Retrieval System*. Prentice-Hall Inc., Eaglewood Cliffs, New Jersy, USA, 1971.

[149] D. Ingham, S. Caughey, and M. Little. Fixing the 'broken-link' problem: the w3objects approach. In *Proceedings of the fifth international World Wide Web conference on Computer networks and ISDN systems*, pages 1255–1268, Paris, France, 1996. Elsevier Science Publishers B. V.

[150] P. Ingwersen. *Information Retrieval Interaction*. Taylor Graham, London, UK, 1992. ISBN 0-947568-54-9.

[151] P. A. Ingwersen. A cognitive view of three selected online search facilities. *Online Review*, 8(5):465–492, 1984.

[152] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing and Management*, 36(2):207–227, 2000.

[153] N. Jardine and v. C. J. The Use of Hierarchical Clustering in Information Retrieval. *Information Storage and Retrieval*, 7:217–240, 1971.

[154] T. Joachims. Optimizing Search Engines using Clickthrough Data. In *KDD '02: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 133–142. ACM Press, 2002.

[155] K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–20, 1972.

[156] K. S. Jones. The Cranfield tests. In K. S. Jones, editor, *Information Retrieval Experiment*, pages 256–284. Butterworths, London, UK, 1981.

[157] K. S. Jones and C. J. van Rijsbergen. Information retrieval test collections. *Journal of Documentation*, 32:63–73, 1976.

[158] G. Kahn, A. Kepner, and J. Pepper. Test, a model-driven application shell. In *Proceedings of the sixth National Conference of the American Association for AI*, pages 814–818, Menlo Park, CA, USA, 1987. AAAI Press.

[159] J. Kamps, C. Monz, M. de Rijke, and B. Sigurbjornsson. Approaches to robust and web retrieval. In E. M. Voorhees and L. P. Buckland, editors, *The Twelfth Text Retrieval Conference(TREC 2003)*, pages 594–600, Gaithersburg, Maryland, USA, 2003. National Institute of Standards and Technology. NIST Special Publication 500-255.

[160] M. Kantrowitz, B. Mohit, and V. Mittal. Stemming and its effects on tf-idf ranking (poster session). In *Proceedings of the twenty-third Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 357–359, New York, NY, USA, 2000. ACM Press.

[161] B. Keller and T. Knutilla. Building an expert diagnostic system. *Ordnance*, pages 44–45, May 1989.

[162] R. B. Kellogg and M. Subhas. Text to hypertext: can clustering solve the problem in digital libraries? In *Proceedings of the First ACM International Conference on Digital Libraries*, pages 144–150, New York, NY, USA, 1996. ACM Press.

[163] D. Kelly, V. D. Dollu, and X. Fu. University of North Carolina's hard track experiments at trec 2004. In E. M. Voorhees and L. P. Buckland, editors, *The Thirteenth Text Retrieval Conference(TREC 2004)*, Gaithersburg, Maryland, USA, 2004. National Institute of Standards and Technology. NIST Special Publication 500-261.

[164] A. Kilgarriff. Corpus Word Usages and Dictionary Word Senses: What is the Match? An Empirical Study. In *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research*, pages 23–39, 1991.

[165] D.-H. Kim and C.-W. Chung. QCluster: Relevance Feedback using Adaptive Clustering for Content-based Image Retrieval. In *SIGMOD '03: Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, pages 599–610, New York, NY, USA, 2003. ACM Press.

[166] B. King. Step-wise Clustering Procedures. *Journal American Statistical Association (JASA)*, 69:86–101, 1967.

[167] D. Knaus, E. Mittendorf, P. Schauble, and P. Sheridan. Highlighting relevant passages for users of the interactive spider retrieval system. In D. K. Harman, editor, *The Fourth text Retrieval Conference (TREC-4)*, pages 233–245,

Gaithersburg, MA, USA, 1995. National Institute of Science technology. NIST Special Publication 500-234.

[168] M. Kobayashi and K. Takeda. Information retrieval on the web. *ACM Comput. Surv.*, 32(2):144–173, 2000.

[169] J. Koenemann. Supporting interactive information retrieval through relevance feedback. In *Conference companion on Human factors in computing systems*, pages 49–50, New York, NY, USA, 1996. ACM Press.

[170] J. Koenemann and N. J. Belkin. A Case for Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 205–212, New York, NY, USA, 1996. ACM Press.

[171] D. H. Kraft, F. E. Petry, W. P. Buckles, and T. Sadasivan. The use of genetic programming to build queries for information retrieval. In *Proceedings of the 1994 IEEE World Congress on Computational Intelligence*, pages 468–473, Piscataway, NJ, USA, 1994. IEEE Press.

[172] R. Krovetz. Homonymy and Polysemy in Information Retrieval. In *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*, pages 72–79. Association for Computational Linguistics, 1997.

[173] R. Krovetz and W. B. Croft. Word Sense Disambiguation using Machine-readable Dictionaries. In *SIGIR '89: Proceedings of the 12th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 127–136. ACM Press, 1989.

[174] R. J. Krovetz. *Word Sense Disambiguation for Large Text Databases*. PhD thesis, University of Massachusetts, 1995.

[175] T. Kurita. An Efficient Agglomerative Clustering Algorithm using a Heap. *Pattern Recognition*, 24(3):205–209, 1991.

[176] K. Kwok, L. Grunfeld, H. Sun, and P. Deng. Trec 2004 robust track experiments using pircs. In E. M. Voorhees and L. P. Buckland, editors, *The Thirteenth Text Retrieval Conference(TREC 2004)*, Gaithersburg, Maryland, USA, 2004. National Institute of Standards and Technology. NIST Special Publication 500-261.

[177] K. L. Kwok, L. Grunfeld, and J. H. Xu. TREC-6 English and Chinese retrieval experiments using PIRCS. In *Text REtrieval Conference*, pages 207–214, Gaithersburg, MA, USA, 1997. National Institute for Science and Technology. NIST Special Publication 500-240.

[178] J. Lafferty and C. Zhai. Document language models, query models and risk minimization for information retrieval. In *Proceedings of the 24th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01)*, pages 111–119, New York, NY, USA, 2001. ACM.

[179] F. W. Lancaster. *Information retrieval Systems: Characteristics, Testing and Evalaution*. John Wiley and Sons, Inc., New York, NY, USA, 1968.

[180] F. W. Lancaster. MEDLARS: Report on the evaluation of its operating efficiency. *American Documentation*, 20(2):119–148, 1969.

[181] L. S. Larkey and W. B. Croft. Combining classifiers in text categorization. In H.-P. Frei, D. Harman, P. Schäuble, and R. Wilkinson, editors, *Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval*, pages 289–297, Zürich, CH, 1996. ACM Press, New York, US.

[182] C. Leacock, G. Towell, and E. Voorhees. Corpus-based Statistical Sense Resolution. In *Proceedings of the ARPA Workshop on Human Language*, pages 260–265, 1993.

[183] D. L. Lee, H. Chuang, and K. Seamons. Document ranking and the vector-space model. *IEEE Software*, 14(2):67–75, 1997.

[184] A. Lelu. Automatic generation of 'hyper-paths' in information retrieval systems: a stochastic and an incremental algorithms. In *Proceedings of the fourteenth annual international ACM SIGIR conference on Research and development in information retrieval*, pages 326–335, New York, NY, USA, 1991. ACM Press.

[185] D. B. Lenat. CYC: A Large-scale Investment in Knowledge Infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.

[186] D. B. Lenat. From 2001 to 2001: Common Sense and the Mind of HAL. In D. G. Stork, editor, *HAL's legacy: 2001's computer as Dream and Reality*. MIT Press, Cambridge, MA, USA, March 1998.

[187] D. B. Lenat, R. V. Guha, K. Pittman, D. Pratt, and M. Shepherd. Cyc: Toward Programs with Common Sense. *Communications of the ACM*, 33(8):30–49, 1990.

[188] M. E. Lesk. Word-word associations in document retrieval systems. *American Documentation*, 20(1):27–38, 1969.

[189] A. Leuski and J. Allen. Improving Interactive Retrieval by Combining Ranked Lists and Clustering. In *Proceedings RIAO 2000 - Content Based Multimedia Information*, pages 665–681. C.I.D, 2000.

[190] C. G. Liungman. *Dictionary of Symbols*. W.W. Norton & Company Ltd, 1994.

[191] R. M. Losee and A. Bookstein. Integrating boolean queries in conjunctive normal form with probabilistic retrieval models. *Information Processing and Management*, 24(3):315–321, 1988.

[192] J. B. Lovins. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11(1-2):11–31, 1968.

[193] A. Lu, M. Ayoub, and J. Dong. Ad hoc experiments using EUREKA. In E. M. Voorhees and D. K. Harman, editors, *The Fifth text Retrieval Conference (TREC-5)*, pages 229–240, Gaithersburg, MA, USA, 1996. National Institute of Science technology. NIST Special Publication 500-238.

[194] G. Lu and B. Williams. An Integrated WWW Image Retrieval System. In *The Fifth Australian World Wide Web Conference, Ausweb'99*, 1999.

[195] Y. Lu, C. Hu, X. Zhu, H. Zhang, and Q. Yang. A Unified Framework for Semantics and Feature Based Relevance Feedback in Image Retrieval Systems. In *MULTIMEDIA '00: Proceedings of the Eighth ACM International Conference on Multimedia*, pages 31–37, New York, NY, USA, 2000. ACM Press.

[196] H. P. Luhn. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2:159–165, 1958.

[197] C. Lundquist, D. A. Grossman, and O. Frieder. Improving relevance feedback in the vector space model. In *Proceedings of the sixth international conference on Information and knowledge management*, pages 16–23, New York, NY, USA, 1997. ACM Press.

[198] T. R. Lynam, C. Buckley, C. L. A. Clarke, and G. V. Cormack. A multi-system analysis of document and term selection for blind feedback. In *CIKM '04: Proceedings of the Thirteenth ACM conference on Information and knowledge management*, pages 261–269, New York, NY, USA, 2004. ACM Press.

[199] M. Magennis and C. J. van Rijsbergen. The potential and actual effectiveness of interactive query expansion. In *SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 324–332, New York, NY, USA, 1997. ACM Press.

[200] S. Mamrak, M. Kaebling, C. Nicholas, and M. Share. A software architecture for supporting the exchange of electronic manuscripts. *Communications of the ACM*, 30(5):408–414, 1987.

[201] S. A. Mamrak, M. J. Kaebling, C. K. Nicholas, and M. Share. Chameleon: A system for solving the data-transportation problem. *IEEE Transactions on Software Engineering*, 15(9):1090–1108, 1989.

[202] E. P. Markatos. On caching search engine query results. *Computer Communications*, 24(2):137–143, 2001.

[203] M. E. Maron and J. L. Kuhns. On relevance, probabilistic indexing and information retrieval. *J. ACM*, 7(3):216–244, 1960.

[204] C. C. Marshall and F. M. Shipman III. Searching for the missing link: Discovering implicit structure in spatial hypertext. In *Hypertext'93: Proceedings of the fifth ACM conference on Hypertext*, pages 217–230, New York, NY, USA, 1993. ACM Press.

[205] C. Meilhac and C. Nastar. Relevance Feedback and Category Search in Image Databases. In *IEEE International Conference on Multimedia Computing and Systems*, 1999.

[206] M. Melucci and N. Orio. A novel method for stemmer generation based on hidden markov models. In *Proceedings of the twelfth international conference on Information and Knowledge Management*, pages 131–138, New York, NY, USA, 2003. ACM Press.

[207] R. Michalski, R. E. Stepp, and E. Diday. Automatic Construction of Classifications: Conceptual Clustering versus Numerical Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5:528–552, 1983.

[208] D. H. Miller, T. Leek, and R. Schwarz. A hidden markov model information retreival system. In *Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, pages 214–221, New York, NY, USA, 1999. ACM Press.

[209] G. A. Miller and W. G. Charles. Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.

[210] M. Mitra, A. Singhal, and C. Buckley. Improving automatic query expansion. In *Proceedings of the 21st annual international ACM SIGIR conference on Re-*

search and development in information retrieval, pages 206–214, New York, NY, USA, 1998. ACM Press.

[211] J. Montgomery, L. Si, J. Callan, and D. A. Evans. Effect of varying number of documents in blind feedback: analysis of the 2003 nrrc ria workshop bfnumdocs experiment suite. In *SIGIR '04: Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 476–477, New York, NY, USA, 2004. ACM Press.

[212] M. Morita and Y. Shinoda. Information Filtering Based on User Behavior Analysis and Best Match Text Retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 272–281, 1994.

[213] J. Morris and G. Hirst. Lexical cohesion computed by theasaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48, 1991.

[214] H. Murray, editor. *The Oxford English Dictionary.* Oxford University Press, Oxford, UK, 1928.

[215] C. Nentwich, L. Capra, W. Emmerich, and A. Finkelstein. xlinkit: a consistency checking and smart link generation service. *ACM Transactions on Internet Technology (TOIT)*, 2(2):151–185, 2002.

[216] G. Newby. Information space based on html structure. In E. Vorhees and D. K. Harman, editors, *TREC-9: The Ninth Text Retrieval Conference*, pages 601–610, Gaithersburg, Maryland, USA, November 2000. National Institute for Science and Technology.

[217] R. Ng and J. Han. Very Large Data Bases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 144–155, Berkeley, CA, 1994. VLDB Endowment.

[218] W. Niblack, R. Barber, and W. Equitz. The QBIC Project: Querying Images by Content Using Color, Texture, and Shape. In *Proceedings of SPIE Electronic Imaging: Science and Technology*, 1993.

[219] H. Nottelmann and N. Fuhr. From retrieval status values to probabilities of relevance for advanced IR applications. *Information Retrieval*, 6(4):363–388, 2003.

[220] D. Nunn, J. Leggett, C. Boyle, and D. Hicks. The rexx project: A case study of automatic hypertext construction. Technical report, Hypertex Research Lab, Texas A&M University, 1988. TAMU 88-021.

[221] M. Orland and R. Saltman, editors. *Manual of Medical Therapeutics*. Little,Bown and Co., Boston, MA, USA, 25 edition, 1986.

[222] C. D. Paice. Another stemmer. *SIGIR Forum*, 24(3):56–61, 1990.

[223] C. D. Paice. An evaluation method for stemming algorithms. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–50, New York, NY, USA, 1994. Springer-Verlag.

[224] C. D. Paice. Method for evaluation of stemming algorithms based on error counting. *Journal of the American Society for Information Science*, 47(8):632–649, 1996.

[225] H. V. D. Parunak. Hypermedia topologies and user navigation. In *Hypertext '89: Proceedings of the Second Annual ACM Conference on Hypertext*, pages 43–50, New York, NY, USA, 1989. ACM Press.

[226] T. Pedersen and R. Bruce. Distinguishing Word Senses in Untagged Text. In *Proc. of the 2nd Conference on Empirical Methods in Natural Language Processing*, pages 197–207, 1997.

[227] T. Pedersen and R. Bruce. Knowledge Lean Word-sense Disambiguation. In *AAAI '98/IAAI '98: Proceedings of the fifteenth national/tenth conference*

on Artificial intelligence/Innovative applications of artificial intelligence, pages 800–805. American Association for Artificial Intelligence, 1998.

[228] G. Ponte and B. Croft. Useg: A retargetable word segmentation procedure for information retrieval. In *Proceedings of the Symposium on Document Analysis and Information Retrieval*, Las Vegas, Nevada, USA, 1996. University of Las Vegas.

[229] J. Ponte. *A language modelling approach to information retrieval*. PhD thesis, University of Massachusetts, Amherst, Pioneer Valley, MA, USA, 1998.

[230] J. Ponte and W. B. Croft. A language modelling approach to information retrieval. In *Proceedings of the 21st ACM SIGIR Conference on Research and Developemnt in Information Retrieval*, pages 275–281, New York, NY, USA, 1998. ACM Press.

[231] M. Porter and V. Galpin. Relevance feedback in a public access catalogue for a research library: Muscat at the Scott Polar Research Institute. *Program*, 22(1):1–20, 1988. See also www.xapian.org.

[232] A. Purandare and T. Pedersen. Senseclusters - Finding Clusters that Represent Word Senses. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence, Sixteenth Conference on Innovative Applications of Artificial Intelligence.*, pages 1030–1031. AAAI Press / The MIT Press, 2004.

[233] Y. Qiu and H.-P. Frei. Concept-based query expansion. In *Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval*, pages 160–169, New York, NY, USA, 1993. ACM Press.

[234] T. Quack, U. Monich, L. Thiele, and B. S. Manjunath. Cortina: a system for large-scale, content-based web image retrieval. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 508–511, New York, NY, USA, 2004. ACM Press.

[235] T. Radecki. Incorporation of relevance feedback into boolean retrieval systems. In *SIGIR '82: Proceedings of the 5th annual ACM conference on Research and development in information retrieval*, pages 133–150, New York, NY, USA, 1982. Springer-Verlag New York, Inc.

[236] T. Radecki. Incorporation of relevance feedback into boolean retrieval systems. In *Proceedings of the 5th annual ACM conference on Research and development in information retrieval*, pages 133–150, New York, NY, USA, 1982. Springer-Verlag New York, Inc.

[237] V. V. Raghavan, P. Bollmann, and G. S. Jung. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems*, 7(3):205–229, 1989.

[238] V. V. Raghavan and S. K. M. Wong. A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science*, 37(5):279–287, 1986.

[239] M. Rasmussen and G. Karypis. gcluto – an interactive clustering, visualization, and analysis system.

[240] D. R. Raymond and F. W. Tompa. Hypertext and the new oxford english dictionary. In *Proceedings of the ACM conference on Hypertext*, pages 143–153, New York, NY, USA, 1987. ACM Press.

[241] T. C. Rearick. *Automating the conversion of text into hypertext*, pages 113–140. McGraw-Hill, Inc., Hightstown, NJ, 1991.

[242] J. T. Rickman. Design considerations for a boolean search system with automatic relevance feedback processing. In *Proceedings of the ACM annual conference*, pages 478–481, New York, NY, USA, 1972. ACM Press.

[243] R. Riner. Automated conversion. In *Hypertext / Hypermedia handbook*, pages 95–111. Intertext Publication, McGraw Hill, Hightstown, New Jersey, USA, 1991.

[244] J. S. Ro. An evaluation of the applicability of ranking algorithms to improve the effectiveness of full-text retrieval. i. on the effectiveness of full-text retrieval. *Journal of the American Society for Information Science*, 39(2):73–78, 1988.

[245] S. E. Robertson. The parametric description of retrieval tests. *Journal of Documentation*, 25:11–27, 1969.

[246] S. E. Robertson. *A Theoretical Model of the Retrieval Characteristics of Information Retrieval Systems*. PhD thesis, University of London, 1976.

[247] S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304, 1977.

[248] S. E. Robertson. On term selection for query expansion. *Journal of Documentation*, 46(4):359–364, 1990.

[249] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146, 1976.

[250] S. E. Robertson, H. Zaragoza, and M. Taylor. Microsoft cambridge at trec-12: Hard track. In E. M. Voorhees and L. P. Buckland, editors, *The Twelfth Text Retrieval Conference(TREC 2003)*, pages 418–426, Gaithersburg, Maryland, USA, 2003. National Institute of Standards and Technology. NIST Special Publication 500-255.

[251] J. J. Rocchio Jnr. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART Retrieval System*. Prentice-Hall Inc., Eaglewood Cliffs, New Jersy, USA, 1971.

[252] J. J. Rocchio Jnr. and G. Salton. Information search optimization and iterative retrieval techniques. In *Proceedings of the AFIPS Fall Joint Computer Conference*, New York, NY, USA, 1965. Spartan Books.

[253] K. Rodden, W. Basalaj, D. Sinclair, and K. Wood. Does Organisation by Similarity Assist Image Browsing? In *CHI '01: Proceedings of the SIGCHI*

conference on Human factors in computing systems, pages 190–197. ACM Press, 2001.

[254] C. Rolker and R. Kramer. Quality of Service Transferred to Information Retrieval: The Adaptive Information Retrieval System. In *Proceedings of the eighth international conference on Information and knowledge management*, pages 399–404, New York, NY, USA, 1999. ACM Press.

[255] W. D. Ross, editor. *Aristotle's Physics: A Revised Text with Introduction and Commentary*. Oxford University Press (OUP), Oxford, England, 1936.

[256] I. Ruthven. Re-examining the potential effectiveness of interactive query expansion. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 213–220, New York, NY, USA, 2003. ACM Press.

[257] I. Ruthven and M. Lalmas. A survey on the use of relevance feedback for information access systems. *Knowledge Engineering Review*, 18(2):95–145, 2003.

[258] W. K. H. Sager and P. C. Lockemann. Classification of ranking algorithms. *International Forum for Information and Documentation*, 1(4):12–25, 1976.

[259] G. Salton. *The SMART retrieval system - Experiments in Automatic Document Processing*. Prentice Hall, Englewood Cliffs, NJ, USA, 1971.

[260] G. Salton. *Automatic Text Processing – The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Series in Computer Science. Addison–Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[261] G. Salton. Developments in Automatic Text Retrieval. *Science*, 253:974–97, 1991.

[262] G. Salton. The SMART document retrieval project. In *Proceedings of the fourteenth annual international ACM SIGIR conference on Research and de-*

velopment in information retrieval, pages 356–358, New York, NY, USA, 1991. ACM Press.

[263] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[264] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297, 1990.

[265] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., Hightstown, NJ, USA, 1986.

[266] G. Salton and A. Wong. Generation and Search of Clustered Files. *ACM Transactions on Database Systems*, 3(4):321–346, 1978.

[267] G. Salton, A. Wong, and C. Yang. A vector space model for information retrieval. *Journal of the American Society for Information Science*, 18(11):613–620, 1975.

[268] G. Salton, C. S. Yang, and C. T. Yu. A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, 26(1):33–44, 1975.

[269] M. Sanderson. Word Sense Disambiguation and Information Retrieval. In *SIGIR '94: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 142–151. Springer-Verlag New York, Inc., 1994.

[270] J. Schachter. del.icio.us - social bookmarks, 2004. http://del.icio.us/.

[271] P. Schäuble. *Multimedia Information Retrieval. Content-Based Information Retrieval from Large Text and Audio Databases*. Kluwer Academic Publishers, Accord Station, MA, USA, 1997.

[272] H. Schütze. Dimensions of Meaning. In *Supercomputing '92: Proceedings of the 1992 ACM/IEEE conference on Supercomputing*, pages 787–796. IEEE Computer Society Press, 1992.

[273] H. Schütze. Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97–123, 1998.

[274] M. M. Sebrechts, J. Cugini, S. J. Laskowski, J. Vasilakis, and M. S. Miller. Visualization of search results: A comparative evaluation of text, 2d, and 3d interfaces. In *Research and Development in Information Retrieval*, pages 3–10, 1999.

[275] J. G. Shanahan, J. Bennett, D. A. Evans, D. A. Hull, and J. Montgomery. Clairvoyance corporation experiments in the trec 2003 high accuracy retrieval from douments (hard) track. In E. M. Voorhees and L. P. Buckland, editors, *The Twelfth Text Retrieval Conference(TREC 2003)*, pages 152–161, Gaithersburg, Maryland, USA, 2003. National Institute of Standards and Technology. NIST Special Publication 500-255.

[276] S. Shatford. Describing a picture: a thousand words is seldom cost effective. *Cataloging & Classification Quarterly*, 4(4):13–30, 1984.

[277] S. Shatford. Analyzing the subject of a picture: a theoretical approach. *Cataloging & Classification Quarterly*, 6(3):39–62, 1986.

[278] D. Shin, S. Nam, and M. Kim. Hypertext construction using statistical and semantic similarity. In *Proceedings of the second ACM international conference on Digital libraries*, pages 57–63, New York, NY, USA, 1997. ACM Press.

[279] M.-L. Shyu, S.-C. Chen, M. Chen, and C. Zhang. Affinity Relation Discovery in Image Database Clustering and Content-based Retrieval. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 372–375, New York, NY, USA, 2004. ACM Press.

[280] M.-L. Shyu, S.-C. Chen, M. Chen, and C. Zhang. A Unified Framework for Image Database Clustering and Content-based Retrieval. In *MMDB '04: Proceedings of the 2nd ACM International Workshop on Multimedia Databases*, pages 19–27, New York, NY, USA, 2004. ACM Press.

[281] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, 1999.

[282] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Research and Development in Information Retrieval*, pages 21–29, New York, NY, USA, 1996. ACM Press.

[283] A. F. Smeaton and P. D. Over. The TREC-2002 video track report. In *TREC 2002 - The Eleventh Text Retrieval Conference*, Gaithersburg, MA, USA, March 2003. National Institute of Science and Technology. NIST Special Publication 500-251.

[284] A. F. Smeaton and C. J. van Rijsbergen. The retrieval effects of query expansion on a feedback document retrieval system. *The Computer Journal*, 26(3):239–246, 1983.

[285] J. R. Smith and S.-F. Chang. Visually Searching the Web for Content. *IEEE MultiMedia*, 4(3):12–20, 1997.

[286] K. Sparck Jones. *Automatic Keyword classification for Information Retrieval*. Butterworths, London, UK, 1971.

[287] K. Sparck Jones. Search term relevance weighting given little relevance information. *Journal of Documentation*, 35(1):30–48, 1979.

[288] K. Sparck Jones, S. Walker, and S. Robertson. A probabilistic model of information retrieval: development and comparative experiments. *Information Processing and Management*, 36:779–808, 2000. Part One.

[289] K. Sparck Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments. *Information Processing and Management*, 36:809–840, 2000. Part Two.

[290] A. Spink, J. Bateman, and B. Jansen. Searching heterogeneous collections on the web: behaviour of excite users. *Information Research*, 4(2), 1998.

[291] A. Spink, B. J. Jansen, and H. C. Ozmultu. Use of query reformulation and relevance feedback by excite users. *Internet Research: Electronic Networking Applications and Policy*, 10(4):317–328, 2000.

[292] A. Spink, D. Wolfram, M. B. J. Jansen, and T. Saracevic. Searching the web: the public and their queries. *J. Am. Soc. Inf. Sci. Technol.*, 52(3):226–234, 2001.

[293] A. Spink and J. Xu. Selected results from a large study of web searching: the excite study. *Information Research*, 6(1), 2000.

[294] M. A. Stricker and M. Orengo. Similarity of Color Images. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 381–392, 1995.

[295] K. Sugiyama, K. Hatano, M. Yoshikawa, and S. Uemura. Refinement of tf-idf schemes for web pages using their hyperlinked neighbouring pages. In *Proceedings of the fourteenth ACM conference on Hypertext and Hypermedia*, pages 198–207, New York, NY, USA, 2003. ACM Press.

[296] L. Sun, J. Zhang, and Y. Sun. Iscas at trec 2004: Hard track. In E. M. Voorhees and L. P. Buckland, editors, *The Thirteenth Text Retrieval Conference(TREC 2004)*, Gaithersburg, Maryland, USA, 2004. National Institute of Standards and Technology. NIST Special Publication 500-261.

[297] J. Surowiecki. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Doubleday, 2004.

[298] M. J. Swain and D. H. Ballard. Color Indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[299] R. C. Swan and J. Allan. Aspect windows, 3-d visualizations, and indirect comparisons of information retrieval systems. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 173–181, New York, NY, USA, 1998. ACM Press.

[300] D. R. Swanson. Some unexplained aspects of the Cranfield tests of indexing performance factors. *Library Quarterly*, 41:223–238, 1971.

[301] J. Tebbutt. User evaluation of automatically generated semantic hypertext links in a heavily used procedural manual. *Information Processing and Management*, 35(1):1–18, 1999.

[302] P. Thistlewaite. Automatic construction and management of large open webs. *Information Processing and Management*, 33(2):161–173, 1997.

[303] E. D. Valle, P. Castagna, and M. Brioschi. Towards a semantic enterprise information portal. In *KCAP'03: Second International Conference on Knowledge Capture (Workshop on Knowledge Management and the Semantic Web)*, New York, NY, USA, October 2003. ACM Press.

[304] C. J. van Rijsbergen. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, 33:106–119, 1977.

[305] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, UK, 2 edition, 1980.

[306] N. Vasconcelos and A. Lippman. Learning from User Feedback in Image Retrieval Systems. In *NIPS'99 - Neural Information Processing Systems*, 1999.

[307] B. C. Vickery. Review of Cleverdon, Mills, and Keen. *Journal of Documentation*, 22:247–49, 1966.

[308] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326, New York, NY, USA, 2004. ACM Press.

[309] E. Voorhees. Overview of the trec 2003 robust retrieval track. In E. M. Voorhees and L. P. Buckland, editors, *The Twelfth Text Retrieval Conference(TREC 2003)*, pages 69–78, Gaithersburg, Maryland, USA, 2003. National Institute of Standards and Technology. NIST Special Publication 500-255.

[310] E. M. Voorhees. Whither music IR evaluation infrastructure: Lessons to be learned from TREC. In *The MIR/MDL Evaluation Project White Paper Collection*, pages 7–13, Champaign, IL, USA, 2002. GSLIS.

[311] S. Walker, S. Robertson, M. Boughanem, G. Jones, and K. S. Jones. Okapi at TREC-6 - automatic ad hoc, vlc, routing, filtering and qsdr. In E. Voorhees and D. Harman, editors, *Proceedings of the 4th Text Retrieval Conference*, pages 125–136, Gaithersburg, MA, USA, 1997. National Institute for Science and Technology. NIST Special Publication 500-240.

[312] R. H. Warren and T. Liu. A review of relevance feedback experiments at the 2003 reliable information access (ria) workshop. In *SIGIR '04: Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 570–571, New York, NY, USA, 2004. ACM Press.

[313] J. Wen, J. Nie, and H. Zhang. Query clustering using user logs. *ACM Trans. Inf. Syst.*, 20(1):59–81, 2002.

[314] J.-R. Wen, J.-Y. Nie, and H.-J. Zhang. Clustering user queries of a search engine. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 162–168, New York, NY, USA, 2001. ACM Press.

[315] G. William, K. W. Church, and D. Yarowsky. A Method for Disambiguating Word Senses in a Large Corpus. *Computers and the Humanities*, 26, 1992.

[316] E. Wilson. Integrated information retrieval for law in a hypertext environment. In *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 663–677, New York, NY, USA, 1988. ACM Press.

[317] J. Xu. *Solving the word mismatch problem through automatic text analysis*. PhD thesis, University of Massachusetts at Amherst, 1997.

[318] J. Xu and W. B. Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, 18(1):79–112, 2000.

[319] J. Xu, J. Zhao, and B. Xu. Nlpr at trec 2004: Robust experiments. In E. M. Voorhees and L. P. Buckland, editors, *The Thirteenth Text Retrieval Conference(TREC 2004)*, Gaithersburg, Maryland, USA, 2004. National Institute of Standards and Technology. NIST Special Publication 500-261.

[320] D. Yarowsky. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of the 33rd conference on Association for Computational Linguistics*, pages 189–196. Association for Computational Linguistics, 1995.

[321] A. Yoshitaka and T. Ichikawa. A Survey on Content-based Retrieval for Multimedia Databases. *IEEE Transactions on Knowledge and Data Engineering*, 11(1):81–93, 1999.

[322] D. Young and B. Schneiderman. A graphical filter/flow model for boolean queries: A prototype implementation and experiment. *Journal of the American Society for Information Science*, 44(6):327–339, 1993.

[323] C. Yu, G. Salton, and C. Buckley. An evaluation of term dependence models in information retrieval. In G. Salton and H. J. Schneider, editors, *Reasearch and development in Information retrieval, Lecture Notes in Computer Science 146*, pages 151–173. Springer-Verlag Inc., New York, NY, USA, 1983.

[324] C. T. Yu, W. S. Luk, and T. Y. Cheung. A statistical model for relevance feedback in information retrieval. *J. ACM*, 23(2):273–286, 1976.

[325] S. Yu, D. Cai, J.-R. Wen, and W.-Y. Ma. Improving pseudo-relevance feedback in web information retrieval using web page segmentation. In *Proceedings of the twelfth international conference on World Wide Web*, pages 11–18, New York, NY, USA, 2003. ACM Press.

[326] U. Zernik. Trainl vs. Trains2: Tagging Word Senses in Corpus. In U. Zernik, editor, *Lexical acquisition: Exploiting on-line resources to build a lexicon*, pages 91–112. Lawrence Erlbaum Associates, London, 1991.

[327] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *CIKM*, pages 403–410, New York, NY, USA, 2001. ACM Press.

[328] T. Zhang, R. Ramakrishnan, and M. Livny. Birch: An Efficient Clustering Method for Very Large Databases. *SIGMOD*, 2:103–114, 1996.