Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Processing and Management 42 (2006) 1137–1150
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Testing the cluster hypothesis in distributed information retrieval

Fabio Crestani [a,*], Shengli Wu [b]

[a] Department of Computer and Information Sciences, University of Strathclyde, Glasgow G1 1XH, UK
[b] School of Computing and Mathematics, University of Ulster, Belfast, UK

Received 24 July 2005; received in revised form 5 December 2005; accepted 5 December 2005
Available online 31 January 2006

Abstract

How to merge and organise query results retrieved from different resources is one of the key issues in distributed information retrieval. Some previous research and experiments suggest that cluster-based document browsing is more effective than a single merged list. Cluster-based retrieval results presentation is based on the cluster hypothesis, which states that documents that cluster together have a similar relevance to a given query. However, while this hypothesis has been demonstrated to hold in classical information retrieval environments, it has never been fully tested in heterogeneous distributed information retrieval environments. Heterogeneous document representations, the presence of document duplicates, and disparate qualities of retrieval results, are major features of an heterogeneous distributed information retrieval environment that might disrupt the effectiveness of the cluster hypothesis. In this paper we report on an experimental investigation into the validity and effectiveness of the cluster hypothesis in highly heterogeneous distributed information retrieval environments. The results show that although clustering is affected by different retrieval results representations and quality, the cluster hypothesis still holds and that generating hierarchical clusters in highly heterogeneous distributed information retrieval environments is still a very effective way of presenting retrieval results to users.

Keywords: Information retrieval; Retrieval results presentation; Clustering; Experimental study; Distributed information retrieval

1. Introduction

With the ever expanding Web, users are faced with a huge number of information resources. How to quickly find what a user needs from this ''information ocean'' is a challenging issue. Although the single database/search engine solution seems to be efficient for that, it may have difficulties to collect all the information needed in practice, especially in relation to resources of the hidden Web. Distributed information retrieval

* Corresponding author. Tel.: +44 141 548 4303; fax: +44 141 548 4523.
E-mail address: f.crestani@cis.strath.ac.uk (F. Crestani).

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2005.12.002

becomes an alternative and viable solution in these cases. In the MIND project,[1] that we have undertaken with four other partners from Europe and USA, we focused on two key issues of distributed information retrieval: resource selection and data fusion. When a user has routine access to a large number of multimedia digital libraries that are globally distributed, the first task he must undertake is selecting some of these resources to which to direct his queries. Considering the large amount of available resources (data sources, collections, Digital Libraries, etc.), this is a tedious work if done manually. So, setting up a broker to accomplish this resource selection task automatically is an attractive solution. The second task the system must undertake for the user is merging, organising and presenting to the user the results retrieved from all the selected resources. This task is called data fusion or results merging. This paper is concerned with this task.

There are several different ways of presenting the merged results to the user. The simplest solution is not to merge the results at all. Results are separately displayed according to their sources. Single-lists are often used for higher usability and effectiveness (Lee, 1997; Shaw & Fox, 1995; Vort & Cotterell, 1999). Another solution is to merge the results based on estimated document relevance. This approach has been investigated by many researchers and is widely used (see Section 2). However, one problem with this approach is that different resources might interpret the query in different ways and producing a single merged list of results is often affected by these different interpretations. Yet another solution is to cluster the results from the different sources based on retrieved document similarity. In this way different interpretations of the query might be captured. In addition, this enables users to browse effectively long results sets (Hearst & Pedersen, 1996). In fact, cluster-based retrieval results browsing is used by several Web search engines such as, for example, Northernlight[2] and Vivisimo.[3] Cluster-based retrieval results presentation is favourable to users since they can see an organised structure of the retrieved set of documents rather than a long list. For example, if a user submits a query with the word ''car'', the system might organise the retrieved documents in several different clusters, such as, for example, ''car rental'', ''auto'', ''buying'', ''sport'', ''car care'', ''classic car'', ''car audio'', etc. (example taken from Vivisimo), which may be helpful for the user to find the information he needs more quickly.

The effectiveness of retrieval results clustering lies in the cluster hypothesis that states that relevant documents tend to be more similar to each other than non-relevant documents (Jardin & van Rijsbergen, 1971). So according to this hypothesis, relevant documents should be found in the same cluster or clusters. It is obvious that the validity and effectiveness of the cluster hypothesis depends on the effectiveness of the clustering. Several hierarchical clustering models have proved to be effective for this purpose in traditional centralised information retrieval environments. However, the validity and effectiveness of the cluster hypothesis has not yet been investigated in distributed information retrieval environments.

Distributed information retrieval (DIR) is different from traditional centralised information retrieval (IR) in several ways:

- In DIR different resources often use different indexing and retrieval models.
- Different resources accessed by a DIR system often include different document collections and there may be overlap between the content of these collections.
- Different resources often have different result lists cut off policies, related to parameters like, for example, document access cost or bandwidth.

Consequently, with respect to clustering the results from different resources, these differences often carry the following consequences:

- The quality of the retrieved results from different resources often vary widely. This is caused by the quality of the search engines used in different resources and/or by the different number of relevant documents retrieved from each resource.

[1] More information about MIND can be found on the project web site at http://www.mind-project.org/.
[2] http://www.northernlight.com/.
[3] http://vivisimo.com/.

- The presentation of results from different resources are often rather different. Some resource may provide titles and/or short automatically generated summaries of the retrieved documents while some others may provide the full text. There might be different reasons for this, often related to design decisions that are local to a specific resource.
- There may be considerable overlap between different resources, with document duplicates being difficult to identify due to different document representations.

So, in general, a DIR environment is much more heterogeneous than a centralised IR environment. In this paper, we present the results of an investigation into the effectiveness of the cluster hypothesis in such an environment. We approach this by investigating the effect of heterogeneity on the clustering. We would like to answer the following research questions:

- Is clustering an effective solution for merging and presenting retrieval results from different and heterogeneous resources?
- Is retrieval results clustering a better retrieval results presentation strategy than single list?
- And finally, does the cluster hypothesis hold in a heterogeneous DIR environment?

The paper is structured as follows: in Section 2 we review some of the related work about results presentation in DIR and the cluster hypothesis. In Section 3 we introduce the methodology and the experimental setting used in our study. Section 4 presents the results under several different settings. Section 5 concludes the paper.

2. Related work

In this section we report on related work on the cluster hypothesis and on retrieval results presentation in IR and DIR.

2.1. The cluster hypothesis

Document clustering has been used in IR for many years. Originally it was used to improve efficiency of search by reducing the number of documents that needed to be compared to the query. The rationale was that by partitioning the collection in clusters, an IR system could restrict the searching to only some of them. It was only with the work of Jardine and van Rijsbergen that clustering became associated with search effectiveness (Jardin & van Rijsbergen, 1971). Also, with the computer becoming faster and faster, clustering lost its use as a way of improving efficiency.

The motivation for the use of clustering as a way of improving retrieval effectiveness lies in the cluster hypothesis. The cluster hypothesis, as proposed by Jardine and van Rijsbergen, states that the association between documents convey information about the "relevance of documents to the request" (Jardin & van Rijsbergen, 1971). In other words, if a document is relevant to an information need expressed in a query, then similar documents are also likely to be relevant to the same information need. So, if similar documents are grouped into clusters, then one of these clusters will contain the relevant documents (or most of them), and the identification of this cluster would improve search effectiveness.

There are various way in which the cluster hypothesis could be exploited. One is by implementing cluster-based retrieval. In cluster-based retrieval the IR system retrieves only one or more clusters in response to the query. These are the clusters whose cluster representation is most similar to the query. However, the actual effectiveness of cluster-based retrieval has yet to be proven. Voorhees (1985) concluded that (single link) hierarchical clustering often fails to group relevant documents in the same clusters, thus making cluster-based retrieval often ineffective. Other studies (see for example (Croft, 1980; Griffiths, Robinson, & Willett, 1984)) failed to prove the actual effectiveness of cluster-based retrieval.

A more promising way of using the cluster hypothesis is in the presentation of retrieval results, that is by presenting in a clustered form only documents that have been retrieved in response to a query. This

query-specific clustering has proved to help users in browsing retrieval results and, in this way, to improve the effectiveness of the IR process.

Hearst and Pedersen studied cluster-based query-specific documents browsing in their Scatter/Gather system. A partitioning clustering algorithm was used for clustering documents retrieved as result of a query. Their work provides evidence validating the cluster hypothesis in this task (Hearst & Pedersen, 1996).

Leuski evaluated document clustering for interactive IR. He used six different hierarchical clustering methods (single link, complete link, group average, weighted average, centroid, and Ward) and obtained the same conclusion as Hearst and Pedersen's, that is that cluster-based retrieved documents browsing improved the effectiveness of interactive IR systems (Leuski, 2001).

Tombros et al. evaluated the effectiveness of query-specific hierarchical clustering with four clustering methods (group average, Ward, complete link, and single link) and five collections (CACM, CISI, LISA, Medline, and WSJ). Their experiments show that, for query-specific clustering, the effectiveness of all clustering methods are very similar, with the exception of single link clustering which performed the poorest (Tombros, Villa, & van Rijsbergen, 2002). These empirical results are somehow in contrast with theoretical work by Everitt that shows that the single link method is theoretically more sound that others (Everitt, 1993). He also showed a number of examples where only single linkage has adequate output, while other methods produced fictitious and non-existent clusters.

Xu and Croft proposed several cluster-based browsing models based on a combination of clustering and language models and re-asserted the effectiveness of cluster-based browsing (Xu & Croft, 1999).

Different from these previous works, the work reported in this paper investigates the effect of heterogeneity on clustering in the DIR environment. More specifically, we focus on the effect on the cluster hypothesis of heterogeneous representations of documents, the presence of document duplicates, and different qualities of retrieval results from different resources, which are key factors in DIR, as Section 2.2 will explain. Our goal is to test if the cluster hypothesis holds also in these extreme conditions and still provides effective query-specific document browsing.

2.2. Results presentation in distributed information retrieval

Distributed Information Retrieval is concerned with retrieving documents from collections that are distributed on different servers and that are retrieved using different IR systems. Various degrees of heterogeneity are possible in a DIR environment. The collection could be distributed on different servers each using the same IR system and producing consistent document scores and representations of retrieval results. On the other hand, collections could be distributed on different servers searchable with radically different IR systems, using different document representations, indexing models, and retrieval results representations.

Key issues of DIR are resource description, resource selection and results merging. These are obviously affected by the level of heterogeneity of the environment. Resource description is concerned with creating a representation of the content of a resource (document collection or set of collections). Resource descriptions are used by the resource selection process, which directs a query to be searched on one or more resources in response to some selection model. Several resource description and selection models have been proposed. Example are: GlOSS/gGlOSS (Gravano, García-Molina, & Tomasic, 1999), CVV (Yuwono & Lee, 1997), CORI (Callan, Lu, & Croft, 1995), decision-theoretic approach (Fuhr, 1999), multi-objective model (Wu & Crestani, 2003), LWP (Hawking & Thistlewaite, 1999), Opt-DocReTrv (Meng et al., 1998), ReDDE (Si & Callan, 2002) and others (Dreilinger & Howe, 1997; Gauch, Wang, & Gomez, 1996). We will not discuss these models here since this paper deals only with results merging.

Results merging is concerned with merging in a single list the results retrieved from the different resources selected by the resource selection process. Several approaches have been proposed. Round-robin is perhaps the simplest merging method, which takes one document in turn from each available result set. Voorhees et al. demonstrated a way of improving the above round-robin method (Voorhees, Gupta, & Johnson-Laird, 1995). By running some training queries, they estimated the performance of each resource. When a new query is encountered, they retrieve a specific number of documents from each resource based on the estimated performance.

Callan et al. proposed a merging strategy based on the scores achieved by both resource and document. The score for each resource is calculated by a resource selection algorithm, CORI, which indicates the "goodness" of a resource with respect to a given query among a set of available resources. The score of a document is the value that the document obtains from that resource, which indicates the "goodness" of that document to a given query among a set of retrieved documents from that resource (Callan et al., 1995).

Calve and Savoy proposed a merging strategy based on logistic regression. Some training queries were needed for setting up the model. Their experiments showed that the logistic regression approach was significantly better than round-robin, raw-score, and normalised raw-score approaches (Calve & Savoy, 2000).

Finally, a quite effective, though not very efficient, results merging method is downloading all the documents estimated to be relevant from different resources and then using a local information retrieval system to re-rank these documents (Lawrence & Giles, 1998). Since downloading the full text of all the documents may affect the efficiency of the process and incur monetary cost on some resources, Tsikrika and Lalmas reported a local search method not using full text but only titles and summary texts as well as resource rank positions for results merging (Tsikrika & Lalmas, 2001).

Merging results into a single list has a number of advantages that are well known in IR. However, when the list is long, past work has shown that it is more effective to present the user with a list of clusters organised in a hierarchical structure. This enables the user to browse the retrieval results quickly and effectively. While several researchers experimented with clustering retrieval results in classical centralised IR obtaining very positive results based on the cluster hypothesis (see Section 2.1), very little work has been done to test if this approach would work in DIR.

The heterogeneity of the DIR environment is the most important issue here. While clustering retrieval results obtained from a single IR system is mainly affected by the choice of the clustering algorithm, clustering in DIR is heavily affected by the retrieval results representation, the IR system used by each resource, the presence of document duplicates, and other factors. These factors may affect the effectiveness of retrieval results presentation and browsing based on the cluster hypothesis, since the clusters obtained in a DIR environment may be considerably different from those obtained in a centralised IR environment and relevant documents may not be grouped together in clusters equally well.

In the rest of this paper we present an experimental investigation into the effects of the heterogeneity of the DIR environment in the clustering of retrieval results and, ultimately, on the validity of the cluster hypothesis in DIR. Section 3 reports the methodology used by the study.

3. Experimental methodology

In this section we briefly present the clustering algorithms and the evaluation measures used in the experimental study. We also outline the methodology upon which the study is based.

3.1. Testing the cluster hypothesis

The purpose of this study is to test if the cluster hypothesis, that has been shown to hold and provide effective cluster-based retrieval results presentation and browsing in classical IR, also holds in DIR.

Since it has been demonstrated that the cluster hypothesis holds for classical IR (see Section 2), that is for an environment in which one or more collections are searched using a single IR system and retrieved documents are clustered based on their full text, we attempt to prove that the cluster hypothesis hold in DIR by comparing effectiveness results between the classical use of the cluster hypothesis and different DIR environments. If we find no significant difference, then we can infer that the cluster hypothesis holds in DIR too. The choice of clustering algorithm, evaluation measures and test collections and IR systems are reported in the following. The experimental method we employ is based on starting with a relatively simple DIR environment and the adding complexity, each time testing if the cluster hypothesis holds. Fig. 1 shows graphically the difference between a classical centralised IR environment and two different DIR environments of varying heterogeneity.

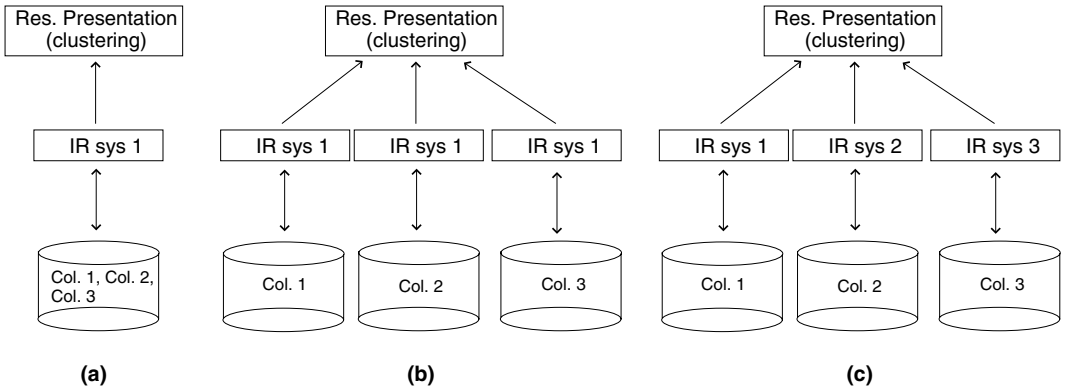

Fig. 1. Different IR and DIR environments: (a) centralised IR; (b) DIR environment with distributed collections indexed and retrieved by the same IR system; (c) DIR environment with distributed collections indexed and retrieved by different IR systems.

3.2. Clustering algorithms

Since the effectiveness of the cluster hypothesis depends on the clustering algorithm used to group documents, the first experimental choice is concerned with selecting one or more clustering methods to use.

There are numerous clustering methods available (Everitt, 1993; Spath, 1980) and many have been used in IR. We may divide them into two main categories: partitioning algorithms and hierarchical algorithms. Hierarchical clustering algorithms can be further divided into two subcategories: divisive methods and agglomerative methods. Among these methods, agglomerative methods, especially the complete link method, are considered by researchers as the most effective methods in terms of retrieval performance (Willett, 1988). Given that the focus of this study is to see if the cluster hypothesis holds for results presentation in DIR, we have to focus our choice of clustering methods to those used for query-specific clustering. In query-specific hierarchical clustering, previous studies found that several agglomerative methods such as complete link, group average, weighted average, and Ward provide very similar performance (Leuski, 2001; Tombros et al., 2002). Since there are no significant differences among these methods, we decided to use only one clustering method, the complete link method. Our choice was rather subjective and not motivated by any theoretical or empirical analysis of which method is best. There seems to be little agreement on which method provides the best results anyway.

We remind the reader that hierarchical agglomerative clustering treats each data point as a singleton cluster, and then successively merges clusters until all points have been merged into a single remaining cluster. In complete link hierarchical clustering the two clusters with the smallest maximum pairwise distance are merged in each step until only one cluster remains (Spath, 1980).

3.3. Evaluation measures

Precision and recall are the main evaluation measures used in IR. However these measures are not applicable to cluster-based retrieval. In order to evaluate the effectiveness of hierarchical clustering structures, an effectiveness measure was proposed by Jardin and van Rijsbergen (1971). This measure is defined as:

$$E = 1 - \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \tag{1}$$

where P and R correspond to the standard measures of precision and recall over the set of documents of a specific cluster, and β is a parameter that reflects the relative importance between precision and recall. Different values of this parameter could be used. For example, if $\beta = 1$, then we assign equal importance to precision and recall; if $\beta = 0.5$, then recall is more important than precision; and if $\beta = 2$, then precision is more important than recall. In the experiments reported in this paper we use $\beta = 1$.

The optimal cluster of a hierarchy for any given query is the cluster that yields the least E value for that query. Therefore, the optimal cluster effectiveness represents the maximum effectiveness that is attainable by a cluster-based search that selects a single cluster in response to each query. This is the way optimal cluster effectiveness is measured in our experiments.

In addition, we need a way to measure the similarity between two different hierarchical clustering structures, to evaluate how close the results of two different hierarchical clusterings are. Dhillon, Fan, and Guan presented a formula for calculating the purity and the entropy of a classification to a standard set of categories (Dhillon, Fan, & Guan, 2001). However, both the classification and the categorisation in their case only had one level. For possible adoption of these measures in our experiments, certain modifications to the original formulae are needed.

Let us see their original formulae first. Suppose we are given c categories and that the clustering algorithm produces k clusters. Cluster π_l's purity is defined as:

$$P(\pi_l) = \frac{1}{n_l} \max_h (n_l^{(h)}) \tag{2}$$

where $n_l = |\pi_l|$ and $n_l^{(h)}$ is the number of documents in π_l that belongs to class h, with $h = 1, 2, \ldots, c$. Note that each cluster may contain samples from different classes. Purity gives the ratio of the dominant class size in the cluster to the cluster size itself. A high purity value implies that the cluster is a "pure" subset of the dominant class.

Entropy is a quality measure, which is defined as follows:

$$H(\pi_l) = -\frac{1}{\log c} \sum_{h=1}^{c} \frac{n_l^{(h)}}{n_l} \log \left(\frac{n_l^{(h)}}{n_l} \right) \tag{3}$$

where c is the number of classes and $n_l^{(h)}$ and n_l are defined as in Eq. (2).

Entropy is a more comprehensive measure than purity. It considers the distribution of classes in a cluster. Note that the above formula always gives normalised entropy values between 0 and 1. An entropy value of 0 means the cluster is comprised entirely of one class, while an entropy value near 1 implies that the cluster contains a uniform mixture of classes.

We want to use the above two measures to evaluate the similarity between two hierarchical clustering structures. In order to do so, we need to adapt these measures designed for flat clustering to hierarchical clustering. Suppose we have a hierarchical structure (binary tree) A to evaluate with regard to another hierarchical structure (binary tree) S, the standard one. We evaluate every inner node $a \in A$ except its root. For every such node, we split S into several non-interleaving subtrees $(s_1, s_2, \ldots, s_m)$. We regard a and $s_1, s_2, \ldots$ as flat clusters, that is, only considering the documents in them but not the structure inside them, so that we are able to use the above method to evaluate the purity of the clustering.

The splitting of S is done as follows. Suppose the inner node $a \in A$ that we want to evaluate includes n documents, then we use a top-down process to split S, which is described by the following algorithm.

Input: a binary tree *S_input* and m (the maximum number of nodes in a subtree)
Output: a list of binary trees *L_output*
Function: This algorithm splits the binary tree *S_input* into a number of subtrees, each of them includes no more than m nodes:

1. put S into a list L;
2. if L is empty then go to step 6; else remove an element E from L;
3. if E includes more than m nodes, then we split E into two subtrees *E_l* and *E_r*, and put *E_l* and *E_r* into list L;
4. if E includes no more than m nodes, we put E into the output list *L_output*;
5. go to step 2;
6. exit.

Table 1
Test collections used in the experiments

Title (years)	Size (MB)	Num. documents	Median num. terms/document	Mean num. terms/document
FT (1991–1994)	564	210,158	316	412.7
WSJ (1987–1989)	267	98,732	245	434.0
FR (1994)	395	55,630	588	644.7

In the above algorithm, we set $2n - 1$ as the value of m for the input. The reason for the use of $2n - 1$ as a threshold is given by the fact that $2n - 1$ is the most probable way to get a set of clusters with an average size of n. When we calculate the purity of each of the inner nodes of A, we average them for the entire clustering structure. We do not include the root node because its purity measure is always 1. A similar way is used for measuring the entropy of a clustering with respect to another clustering. Both purity and entropy are in their normalised form, taking values in $[0, 1]$. However, for purity, a value close to 1 means that the tested classification is similar to the reference clustering, while it the opposite for entropy.

3.4. Retrieval systems and test collections

In order to be able to generalise the results of the experimental study, we decided to use different test collections and different IR systems. This enables us to simulate an heterogeneous DIR environment.

Three full-text collections were used in the experiments. They were: the Financial Times (FT) 1991–1994, the Wall Street Journal (WSJ) 1987–1989, and the Federal Register (FR) 1994. These are standard test collections used in TREC (Voorhees & Harman, 1996) and have been used by many researchers. Table 1 summarises the characteristics of these collections. For the evaluation a set of 50 queries was used, corresponding to TREC topics 251–300. All results reported are averaged over the entire set of queries.

Three information retrieval systems were used in the experiment: Glass (G),[4] Smart (S) (Salton, 1990; Sumner & Shaw, 1996) and Lemur (L) (Ogilvie & Callan, 2001).

In order to simulate an heterogeneous DIR environment, we assumed the existence of three different resources, each with a different collection and a different IR system. In the experiments reported in Section 4.4, Smart was used to retrieve from the WSJ collection, Glass from the FT collection and Lemur from the FR collection. This setting is supposed to simulate a real DIR environment used by news professionals, where resources are distributed on completely different and independent Digital Libraries.

4. Experimental results and analysis

In this section we report the experimental results of the study and attempt an analysis. In Section 4.1 we report the results of the testing of the cluster hypothesis for results presentation in a DIR environment of type b of Fig. 1, in which the IR system produces the same representations of the retrieved documents across all collections. We varied the retrieved document representations from title only to full text. In Section 4.2 we simulate the situation in which the same IR system produces a different representation of the retrieval results for each collection. This still corresponds to type b of Fig. 1, but with the added complexity of the heterogeneous representation of the retrieval results. In Section 4.3 we increase the complexity even more by introducing the presence of duplicates in the retrieved documents sets, where duplicates are represented in different ways. Finally, in Section 4.4 we add the effect of different and varying retrieval qualities of the IR systems used by the different resources. This last simulated DIR environment, which corresponds to type c of Fig. 1, is the most heterogeneous DIR environment we could think of. This kind of environment is also similar to real life DIR applications and corresponds to the kind of environments we experimented with in the EU projects MIND and PENG.[5]

[4] Glass is an experimental IR system designed and developed by Mark Sanderson. No reference is available for this system.
[5] More information about the PENG project can be found at http://www.peng-project.net/.

4.1. Homogeneous retrieved document representations

We first experimented with homogeneous retrieved document representations. This corresponds to a DIR environment in which the same IR system is used by all the different resources and the same retrieved document representation is produced by each resource. This is not dissimilar to having all collections in the same resource and retrieved using the same IR system, since producing a combined ranking of the retrieved results is easy given that documents are assigned directly comparable scores. So, there is basically no difference between this type of environment and a classical centralised IR environment. What we tested was the effect that different retrieved document representations have on the clustering, when the same representation format was used for each collection. So, we assumed that the IR system would retrieve for each of the top 100, 150 and 200 documents in the retrieved list one of the following:

- only the title;
- the title and the first 50 words;
- the title and the first 100 words;
- the title and the first 150 words;
- the full text.

We then clustered the documents in the retrieved sets and compared the different clusterings obtained, in particular comparing them with the clustering obtained from the full text of the documents. Table 2 shows the average effectiveness obtained using the G system over the 50 queries. Results obtained using the S and L systems are very similar and are not reported here due to space limitations. The table shows that the effectiveness figures of the different hierarchies produced with the different document representations are not very different from those obtained using the full text. Statistical significance tests (two tailed t test with 2% level of significance) show that only the figures in bold are statistically significantly different from the figures obtained with full text. This indicates that, with the exception of a few cases involving very short representations of the retrieved documents, the effectiveness of the clustering is basically the same whatever document representation is used. Tables 3 and 4, however, show that the hierarchical clusterings produced using these different

Table 2
Average effectiveness of clusterings with homogeneous document representations

Num. documents	Title only	Title + 50 words	Title + 100 words	Title + 150 words	Full text
100	0.709	**0.715**	0.713	0.699	0.660
150	0.707	0.708	0.707	0.689	0.664
200	**0.715**	0.684	0.696	0.680	0.638

Table 3
Average purity of hierarchies with homogeneous document representations compared to clusterings with full text

Num. documents	Title only	Title + 50 words	Title + 100 words	Title + 150 words	Full text
100	0.587	0.664	0.698	0.743	0.769
150	0.569	0.646	0.690	0.721	0.752
200	0.557	0.638	0.677	0.714	0.744

Table 4
Average entropy of clusterings with homogeneous document representations compared to clustering with full text

Num. documents	Title only	Title + 50 words	Title + 100 words	Title + 150 words	Full text
100	0.298	0.243	0.223	0.192	0.176
150	0.282	0.234	0.206	0.188	0.169
200	0.272	0.224	0.200	0.180	0.163

document representations are very different. The purity and entropy figures obtained using short document representations are rather different from those obtained using long document representations (full text and title and 150 words). Figures reported in Tables 3 and 4 are all statistically significantly different from those obtained using full text.

The effectiveness results and the purity and entropy results put together enable us to conclude that although the clusterings obtained with different document representations are very different, they enable to achieve the same level of effectiveness, when document representations are sufficiently long. In other words, most retrieved documents end up in different clusters depending on the document representation employed, but relevant documents tend to end up together in the same clusters nonetheless. This suggests that the cluster hypothesis holds across a variety of retrieved document representations.

4.2. Heterogeneous retrieved document representations

Homogeneity among retrieved document representations in a DIR environment is not common. The most common case is that different resources produce different retrieved document representations. So we decided to test the validity of the cluster hypothesis in the presence of heterogeneity in the retrieved document representations. In order to isolate the effect of heterogeneity, we still assumed that each resource used the same IR system, but that each resource produced a different document representation, ranging among the possibilities reported in the previous section. Since there could be a large number of possibilities of heterogeneous conditions, we made up a few cases and experimented in these conditions only. Tables 6–8 report the average effectiveness, purity and entropy obtained for the eight cases of heterogeneous retrieved document representations in the retrieved set reported in Table 5. Case h1, for example, refers to the case in which IR system 1 retrieves only the title of documents estimated to be relevant, the same for IR system 2, while IR system 3 retrieves the title and the first 50 words. Since we do not carry out any resource selection, each resource contributes approximately the same number of documents to the combined retrieval set (35% for IR systems 1 and 2 and 30% for IR system 3, for simplicity). While we experimented with a large number of combinations, the above 8 seem good examples of how heterogeneous the situation could be. They range from very short representations of the retrieved documents (h1 and h2) to increasingly long document representations (h7 is full text for all systems and can be considered the benchmark), with one very mixed case (h8).

Tables 6–8 report the results obtained with different sizes of the set of retrieved documents produced by the S system. Very similar results were obtained with the L and G systems.

Table 6 shows the average effectiveness of hierarchical clustering obtained in the eight conditions indicated above. Only the figure in bold is statistically significantly different from the effectiveness obtained using the full text for all documents (case h7). This shows that, despite the great variety of heterogeneity in retrieved document representations, there is no significant difference between the effectiveness obtained by the different clustering.

Tables 7 and 8 confirm what we have already seen in Section 4.1, that is that the clusterings obtained in different conditions are very different from those obtained using the full text, as the low values of the average entropy and the high values of the average purity show.

Table 5
Cases of heterogeneous representations in the retrieved document set used in the experiments

Case	35%	35%	30%
h1	Title	Title	Title + 50 words
h2	Title	Title + 50 words	Title + 100 words
h3	Title + 50 words	Title + 100 words	Title + 150 words
h4	Title + 100 words	Title + 150 words	Title + 200 words
h5	Title + 150 words	Title + 200 words	Title + full text
h6	Title + 200 words	Title + full text	Title + full text
h7	Title + full text	Title + full text	Title + full text
h8	Title	Title + 100 words	Title + full text

Table 6
Average effectiveness of hierarchies with heterogeneous document representations

Num. documents	h1	h2	h3	h4	h5	h6	h7	h8
100	0.698	0.706	0.687	0.661	0.655	0.639	0.669	0.701
150	0.701	0.708	0.690	0.681	0.644	0.656	0.693	0.680
200	**0.715**	0.704	0.686	0.672	0.673	0.665	0.687	0.704

Table 7
Average purity of clusterings with heterogeneous document representations compared to clustering with full text

Num. documents	h1	h2	h3	h4	h5	h6	h7	h8
100	0.606	0.603	0.675	0.614	0.734	0.765	0.644	0.621
150	0.587	0.579	0.660	0.698	0.721	0.748	0.640	0.604
200	0.571	0.563	0.645	0.677	0.702	0.731	0.623	0.586

Table 8
Average entropy of clusterings with heterogeneous document representations compared to clustering with full text

Num. documents	h1	h2	h3	h4	h5	h6	h7	h8
100	0.284	0.286	0.234	0.206	0.194	0.174	0.260	0.275
150	0.270	0.273	0.220	0.196	0.181	0.167	0.237	0.261
200	0.263	0.267	0.214	0.195	0.180	0.167	0.231	0.254

Again, even in the presence of highly heterogeneous retrieved document representations, the cluster hypothesis seem to hold, since relevant documents tend to end up clustered together nonetheless.

4.3. Document duplicates in the retrieved document sets

A problem that often occurs in DIR is related to the presence of duplicates in the retrieved document sets. This is due to the fact that different resources might have the same collections or there might be some level of overlap in the documents contained in their collections. Document duplicates are often difficult to identify given that different resources might use different document identifications. This is not a serious problem when all retrieved documents are represented in the same way (i.e. in homogeneous document representation conditions), since any decent document similarity function should be able to identify duplicates and place them in the same clusters. However, it is a problem with heterogeneous document representations, since the same document could be represented in totally different ways in different retrieved documents sets. We therefore decided to test the effectiveness of the cluster hypothesis in the presence of duplicates in the retrieved sets.

Table 9 shows the average effectiveness of the clustering obtained in the eight different conditions of heterogeneous retrieved document representations reported in Section 4.2. The values were obtained using the G IR system with 10, 20, and 40 duplicates in the 200 retrieved documents. The only value that is significantly different from the average effectiveness obtained using the full text of retrieved documents and no duplicates is the one reported in bold, which refers to the case of very short document representations. This is very likely

Table 9
Average effectiveness of clusterings with heterogeneous document representations and duplicates

Num. duplicates	h1	h2	h3	h4	h5	h6	h7	h8
10	0.694	0.679	0.664	0.663	0.667	0.636	0.642	0.657
20	0.700	0.713	0.686	0.683	0.669	0.638	0.698	0.704
40	**0.631**	0.639	0.611	0.601	0.590	0.593	0.607	0.638

Table 10
Average purity of clusterings with heterogeneous document representations and duplicates compared to clustering with full text

Num. duplicates	h1	h2	h3	h4	h5	h6	h7	h8
10	0.577	0.570	0.669	0.708	0.742	0.784	0.657	0.621
20	0.603	0.597	0.691	0.727	0.758	0.797	0.674	0.639
40	0.659	0.644	0.740	0.769	0.792	0.827	0.721	0.677

Table 11
Average entropy of clusterings with heterogeneous document representations and duplicates compared to clustering with full text

Num. duplicates	h1	h2	h3	h4	h5	h6	h7	h8
10	0.263	0.267	0.208	0.185	0.166	0.143	0.216	0.235
20	0.251	0.253	0.199	0.178	0.159	0.137	0.209	0.228
40	0.225	0.232	0.175	0.157	0.143	0.122	0.185	0.210

due to the small amount of information about the document content available to the clustering algorithm to use for computing document similarities. No other value is significantly different. Even slightly better values and still no significantly different from those obtained with full text and no duplicates were obtained using the S and L systems.

The values of average purity and entropy reported in Tables 10 and 11 are very similar to those reported in Tables 7 and 8 and again prove that clusterings obtained from full text and no duplicates are very different from those obtained from heterogeneous document representations and duplicates. Yet, the cluster hypothesis seems to hold and duplicates of relevant documents are being placed in the clusters with other relevant documents despite their different representations.

4.4. Heterogeneous retrieval qualities

Finally, we tested the effect of different retrieval qualities. This situation is typical of a DIR environment in which each resource uses a different IR system, with different systems having different effectiveness levels and retrieving different numbers of relevant documents. In this case it becomes necessary to employ some form of resource selection (based on some form of resource description), and results merging. Since the purpose of this paper is not to investigate resource description, selection or results merging, but to test the effectiveness of the cluster hypothesis in a DIR environment, we decided to assume the worst case scenario for the DIR environment, that is complete ignorance of the resources and of the performance of their IR systems. In this situation we do not have resource descriptions and resource selection is not possible, so the best strategy is to retrieve the same number of documents from each resource. In addition, the only results merging strategy that can be used is round-robin, since we have no information to compare document scores. So, we compared the effectiveness of a single list obtained using round-robin with hierarchical clustering. Different sizes of the combined set of retrieved documents and different, but homogeneous, retrieved document representations were used.

In order to compare the effectiveness of the single-lists with that of hierarchical clusterings we treated the single list of n documents as n clusters, including the top $1, 2, \ldots, n$ documents, respectively. Then we used Eq. (1) to compute the effectiveness of each cluster and the effectiveness of the single-list is the maximum we obtain from all clusters. In such a way, a single-list and a hierarchical structure become comparable.

Table 12
Average effectiveness of clusterings with different retrieval qualities

Num. documents	Round-robin (single-list)	Title only	Title + 50 words	Title + 100 words	Title + 150 words	Full text
100	0.698	0.695	**0.671**	**0.662**	**0.652**	**0.660**
150	0.699	0.710	0.696	**0.648**	**0.680**	**0.650**
200	0.697	0.711	**0.680**	**0.667**	**0.664**	**0.669**

Table 12 reports the average effectiveness results. Figures in bold are statistically significantly different from the single list figures. The table shows that clustering produces better effectiveness value, on average, for all retrieved document representation, except in the case of very short representations. These results show that clustering is a better strategy than round-robin in case of poor or unknown resource descriptions and prove that the cluster hypothesis is an effective way of presenting retrieval results in even the most heterogeneous DIR environment.

5. Conclusions

In this paper we reported on an experimental study on query-specific document clustering with various kinds of representation, duplicated documents, and results with disparate qualities, which frequently happen in the distributed information retrieval environment. The aim of our study was to test if the cluster hypothesis would hold in a DIR environment and if presenting retrieved documents in a hierarchical clustering would still be as effective as it is in classical IR.

Our results show that the cluster hypothesis does hold in even the most heterogeneous and complex DIR environments and that it fails only when document representations are very short and there is not sufficient information for the similarity function used by the clustering to group relevant documents together. It remains to be proved if more complex short document representations, like for example document summaries, would enable to improve the situation. Nevertheless, our results suggest that, despite heterogeneous document representations begin different from full text, duplicates in the retrieved set and varying retrieval quality, clustering retrieval results for presentation to users in a DIR environment is a very good alternative to merging the results in a single list. Our results show that presenting results in a hierarchical clustering is as good a result presentation strategy in DIR as it is in classical IR. Indeed, one could argue that hierarchical clustering of retrieval results is a better result presentation strategy than a single merged list in DIR as it is in classical IR, thanks to the cluster hypothesis.

Acknowledgements

This work was supported by the European Commission under the Information Society and Technology Project MIND (IST-2000–26061). More information about MIND can be found at http://www.mind-project.org/.

References

Callan, J. K., Lu, Z., & Croft, W. (1995). Searching distributed collections with inference networks. In *Proceedings of the 18th annual international ACM SIGIR conference, Seattle, USA* (pp. 21–28).

Calve, A. L., & Savoy, J. (2000). Database merging strategy based on logistic regression. *Information Processing and Management, 36*(3), 341–359.

Croft, W. B. (1980). A model of cluster searching based on classification. *Information Systems, 5*(3), 189–195.

Dhillon, I. S., Fan, J., & Guan, Y. (2001). Efficient clustering of very large document collections. In *Data mining for scientific and engineering applications*. Kluwer Academic Publishers.

Dreilinger, D., & Howe, A. (1997). Experiences with selecting search engines using metasearch. *ACM Transactions on Information Systems, 15*(3), 195–222.

Everitt, B. (1993). *Cluster analysis* (3rd ed.). London, UK: Edward Arnold.

Fuhr, N. (1999). A decision-theoretic approach to database selection in networked IR. *ACM Transactions on Information Systems, 17*(3), 229–249.

Gauch, S., Wang, G., & Gomez, M. (1996). Profusion: Intelligent fusion from multiple distributed search engines. *Journal of Universal Computer Science, 2*(9), 637–649.

Gravano, L., García-Molina, H., & Tomasic, A. (1999). Gloss: text-source discovery over the internet. *ACM Transactions on Database Systems, 24*(2), 229–264.

Griffiths, A., Robinson, A., & Willett, L. A. (1984). Hierarchical agglomerative clustering methods for automatic document classification. *Journal of Documentation, 40*(3), 175–205.

Hawking, D., & Thistlewaite, P. (1999). Methods for information server selection. *ACM Transactions on Information Systems, 17*(1), 40–76.

Hearst, M. A., & Pedersen, J. O. (1996). Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of the 19th annual international ACM SIGIR conference, ZTH, Zürich, Switzerland* (pp. 76–84).

Jardin, N., & van Rijsbergen, C. J. (1971). The use of hierarchical clustering in information retrieval. *Information Storage and Retrieval, 7*(5), 217–240.

Lawrence, S., & Giles, C. L. (1998). Searching the world wide web. *Science, 280*(3), 98–100.

Lee, J. H. (1997). Analysis of multiple evidence combination. In *Proceedings of the 20th annual international ACM SIGIR conference, Philadelphia, PA, USA* (pp. 267–275).

Leuski, A. (2001). Evaluating document clustering for interactive information retrieval. In *Proceedings of ACM international conference on knowledge management, Atlanta, GA, USA* (pp. 33–40).

Meng, W., Liu, K., Yu, C., Wang, X., Chang, Y., & Rishe, N. (1998). Determining text databases to search in the internet. In *Proceedings of the 24th international conference on very large data bases, New York City, USA* (pp. 14–25).

Ogilvie, P., & Callan, J. (2001). Experiments using the lemur toolkit. In *Proceedings of the 2001 text retrieval conference, Gaithersburg, MD, USA* (pp. 103–108).

Salton, G. (1990). Full text information processing using the smart system. *Data Engineering Bulletin, 13*(1), 2–9.

Shaw, J. A., & Fox, E. A. (1995). In *Proceedings of 3rd text retrieval conference (trec-3), Gaithersburg, MD, USA*.

Si, L., & Callan, J. (2002). Using sampled data and regression to merge search engine results. In *Proceedings of the 25th annual international ACM SIGIR conference, Tempere, Finland* (pp. 19–26).

Spath, H. (1980). *Cluster analysis algorithms for data reduction and classification of objects*. Ellis Horwood Limited.

Sumner, R. G., Jr., & Shaw, W. M., Jr. (1996). *An investigation of relevance feedback using adaptive linear and probabilistic models, Gaithersburg, MD, USA, November 20–22*.

Tombros, A., Villa, R., & van Rijsbergen, C. J. (2002). The effectiveness of query-specific hierarchic clustering in information retrieval. *Information Processing and Management, 38*(4), 559–582.

Tsikrika, T., & Lalmas, M. (2001). Merging techniques for performing data fusion on the web. In *Proceedings of the 2001 ACM CIKM international conference on information and knowledge management, Atlanta, GA, USA* (pp. 127–134).

Voorhees, E. M. (1985). The effectiveness and efficiency of agglomerative hierarchical clustering in document retrieval (Tech. Rep. No. TR 85-705). Ph.D. Thesis, Department of Computer Science, Cornell University, Ithaca, NW, USA.

Voorhees, E. M., Gupta, N. K., & Johnson-Laird, B. (1995). Learning collection fusion strategies. In *Proceedings of the 18th annual international ACM SIGIR conference, Seattle, WA, USA* (pp. 172–179).

Voorhees, E. M., & Harman, D. K. (Eds.). (1996). *Proceedings of the 5th text retrieval conference, Gaithersburg, MD, USA*.

Vort, C. C., & Cotterell, G. A. (1999). A fusion via a linear combination of scores. *Information Retrieval, 1*(3), 151–173.

Willett, P. (1988). Recent trends in hierarchical document clustering: A critical review. *Information Processing and Management, 24*(5), 557–597.

Wu, S., & Crestani, F. (2003). Distributed information retrieval: A multiobjective resource selection approach. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 11*, 83–100.

Xu, J., & Croft, W. B. (1999). Cluster-based language models for distributed retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference, Berkeley, CA, USA* (pp. 254–261).

Yuwono, B., & Lee, D. (1997). Server ranking for distributed test retrieval systems on the internet. In *Proceedings of the fifth international conference on database systems for advanced application, Melbourne, Australia* (pp. 41–50).