Mapping physical artifacts to their Web counterparts: A case study [*]

xxxxxxx xxxxxxx   xxxxxxx xxxxxxx
xxxxxxx xxxxxxx xxxxxxx, xxxxxxx xxxxxxx xxxxxxx, xxxxxxx
{xxxxxxx, xxxxxxx}@xxxxxxx.xxxxxxx

ABSTRACT

Many kinds of documents—newspapers, books, product catalogs, directories, etc—exist in both a physical (paper) and virtual (Web) form. Few approaches to knowledgment management and digital libraries fully exploit the opportunities afforded by this fact. Motivated by the goal of seamless integration of physical artifacts and their Web counterparts, we describe a large-scale case study of one aspect of this relationship. Based on a corpus of hundreds of real-world product catalogs, we measure the effectiveness of hand-held scanner/OCR devices for the task of automatically retrieving a catalog's authoritative Web counterpart (the vendor's home page). We find that, despite OCR errors, text fragments scanned from product catalogs can serve as reasonably effective queries for retrieving the Web counterparts. Furthermore, the effectiveness of the technique increases with multiple scanned text fragments. Our main technical contribution is a novel machine learning approach to adaptively merging the retrieved documents from multiple scans.

1. INTRODUCTION

Many kinds of documents—newspapers, books, product catalogs, directories, invoices, resumes, advertisements, etc—have both a physical (ie, paper) manifestation as well as a virtual (Web) form. The full text content of the newspaper that you read this morning is probably available at the newspaper's Web site; the transactions listed on the bank statement you received yesterday are probably available through your bank's Web site.

Few approaches to knowledgment management and digital libraries take advantage of the opportunities afforded by this observation. For example, suppose you fill in an ordering form in a paper product catalog, and the catalog's vendor has a Web order form. What technology could enable one's writing on a piece of paper to be automatically converted to the submission of the Web form?

We are motivated by the goal of enriching the integration between physical artifacts and their Web counterparts. In this paper, we focus on one specific aspect of this relationship: developing technologies to identify the authoritative Web resource to which some specific physical artifact corresponds. Given the ubiquity of electronically produced documents, and the resilience of the "paper-full office", we anticipate that this sort of enabling technology is relevant to a wide variety of knowledge management and digital library applications.

To investigate these issues, we carried out a case study on a corpus of hundreds of real-world product catalogs. We assess the effectiveness of off-the-shelf hand-held pen scanner/OCR devices for the task of automatically retrieving a catalog's authoritative Web counterpart (the vendor's home page). We find that, despite OCR errors, text fragments scanned from product catalogs can serve as reasonably effective queries for retrieving the Web counterparts.

The remainder of this paper is organised as follows. First, we describe our catalog retrieval task and data-set in more detail (Sec. 2). We then describe experiments for two scanning scenarios: first, we consider the effectiveness of individual scanned text fragments in isolation (Sec. 3); second, we describe a novel machine-learning approach to adaptively merging the results from multiple scanned fragments (Sec. 4). We conclude with a discussion of related work and future directions (Secs. 5–6).

2. THE CATALOG RETRIEVAL TASK

As depicted in Fig. 1, the specific task we address is that of identifying the authoritative Web document associated with a given physical paper product catalog. Specifically, we assume that the reader has used a pen scanning device to scan a number of text fragments from the paper catalog, and she seeks to find the Web home page of the corresponding company.

We seek to minimise the reader's effort along two dimensions: the number of text fragments that must be scanned in order to locate the desired Web document, and the number of candidate documents that the reader must examine. Our experiments demonstrate that these two costs trade off against each other: if the user is willing to scan just one or two fragments, (s)he will have to wade through a large num-

[*] Submitted to the Mobile & Ubiquitous Information Access workshop, Mobile HCI 2004

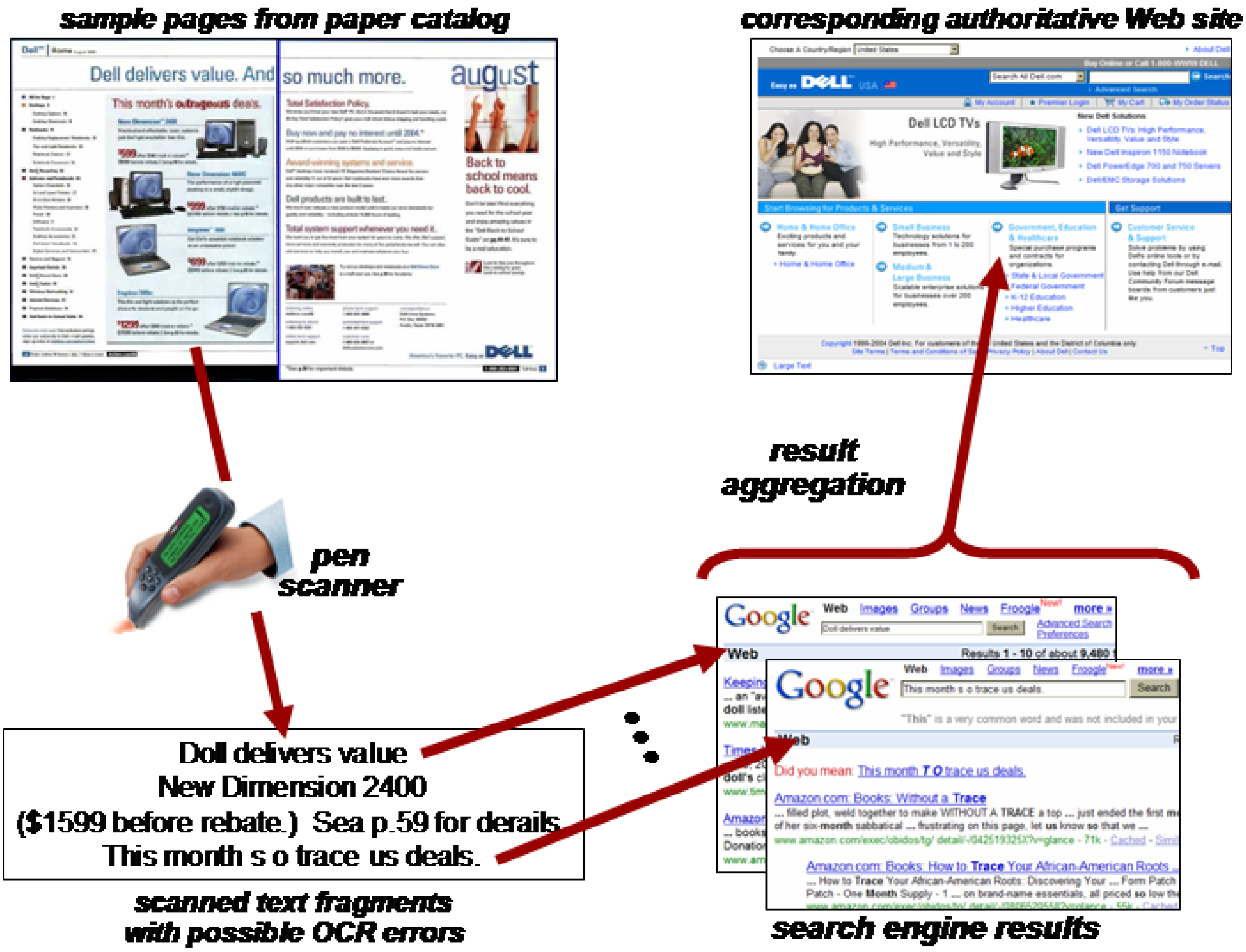
sample pages from paper catalog
corresponding authoritative Web site
result aggregation
pen scanner
Doll delivers value
New Dimension 2400
($1599 before rebate.) Sea p.59 for derails
This month s o trace us deals.
scanned text fragments
with possible OCR errors
search engine results

Figure 1: Sample pages from a product catalog (upper-left), some scanned text fragments (lower-left), and the authoritative Web document with which the catalog is associated (upper-right).

ber of potential documents before finding the right one. On the other hand, if she is willing to invest the effort scanning additional fragments, then the correct document will appear near the top of the list of suggestions.

To explore these ideas, we constructed a dataset from the scanned catalogs available from `catalogs.google.com`. We picked catalogs randomly, and made the assumption that OCR errors of a catalog would be independent from how likely that catalog was to have a website, or how easily retrievable that website would be. We downloaded the first ten scanned pages from each selected catalog, and attempted to perform OCR on them using a commercial OCR program. If the catalog passed a threshold of OCR readability then we accepted it and downloaded and performed OCR on the remaining pages. If the catalog failed to pass the threshold then it was rejected and another was selected. This process was followed until we had selected 295 catalogs from 2000 attempts.

The resulting text from the OCR process was then used as the basis for simulating the actions of a pen scanning device on the original catalog. As the page images on `catalogs.google.com` are scanned from the original paper catalogs, we simulate the use of a pen scanner by running OCR over these images. Assuming equal sophistication of OCR engines, we expect to get similar OCR mistakes with the original paper media and an actual pen scanner.

As described in detail below, the OCR text was processed to generate simulated scan phrases. Each phrase consisted of at least two words, possibly with OCR errors, representing the text which would have been retrieved directly from a pen scanner from the original catalog. Each scan phrase is sent to a Web search engine as a query; we used Google in our experiments. Each phrase was submitted in two ways: as an unquoted bag of terms, and as a quoted sequence of terms. The top 20 results from each query were recorded.

Additionally for each catalog, Google provided the catalog title, and in 96% of the cases the URL of the catalog's homepage. In the remaining 14 cases where the URL was missing, we manually searched the Web to find the authoritative Web site. In 4 of these cases, the catalog appeared to have no on-line version; these catalogs were kept in the collection, but our experiments below will always fail to locate their Web sites.

To enable large-scale experimentation, we simulated a user scanning text from the original paper catalogs. This way we automatically generate a large number of potential queries, each with any associated scanning errors which might have been also introduced via use of a pen scanner, and by selecting parts of the OCR text which are prominant, we aim to pick similar pieces of text as a user would do. To create a simulated scan for a catalog, we first randomly select one of the pages, and then select fragments of OCR text in one of three ways:

ALLCAP: The candidate scans are the sequences of two or more entirely capitalised words, excluding terms in a list of 36 stopwords.

INITCAP: Same as ALLCAP, with only the first letter of each capitalised.

1STLINE: The candidates are the first sentences (up to a maximum of 10 words) of any paragraph of text containing at least 20 words.

One drawback is that since we do not employ any form of semantic analysis, we have no way of choosing between meaningful phrases and those which, although prominant, are in themselves semanticly detached from the topic of the catalog (eg, "Table of Contents"). Therefore, we also generate additional scans as follows:

TITLE: A scan for the title of each catalog, if available.

HAND: Manually generated scans.

The titles are not taken from the OCR text output, but rather from the additional data provided by Google. As such, TITLE scans have a large advantage over the other types: they do not suffer from random OCR errors, or complete OCR failure (due, for example, to a title being set in unusual display font). It should also be noted that many titles were also found in the HAND scans.

Informally, a random sample of 100 title pages of catalogs from the entire collection on Google, only 15% were found to have a scannable title. A fairer analysis of our specific catalog collection showed that in only 59% of the cases did the TITLE scan appear anywhere in the OCR output of the entire catalog. While our experiments show that simulated TITLE scans are highly effective, we do not believe these results are useful in practice due to the difficulty of retrieving them with a pen scanner.

The HAND scans were selected manually with the intent of helping to identify the original catalog. They were selected by looking at the image of a randomly selected page from the catalog, with the OCR output to the side. A piece of text was selected from the image and the corresponding part of the OCR text was input. Our experiments show that HAND scans are more effective than the automatically generated versions, not suffering from the problem where selected scans have no semantic relevance. On the other hand, HAND scans also tend to suffer more from OCR errors for their length: bad OCR output will confuse the automatic scan selection algorithms, thus they tend not to be accepted as readily. HAND scans were only generated for 50 of the 300 catalogs.

	ALLCAP	INITCAP	1STLINE	HAND
Error rate	14%	19%	32%	28%

Figure 2: OCR error rates of the four scan types. TITLEs do not have OCR errors, but do not occur in 41% of the catalogs.

	Academy Chicago Publishers `www.academychicago.com`	The Apothecary `www.the-apothecary.com`
ALLCAP	ACADEIYIY CHICAGO	ESSENTIAL FATTY ACIDS
INITCAP	New York	Biotec Foods
1STLINE	His first book Pilgrimage Tales from the Open Road published	1 We ve chosen to devote this catalog to a
TITLE	academy chicago publishers	the apothecary
HAND	a unique volume from wicker park press	the apothecary order certificate

Figure 3: Actual examples of the five scan types for two catalogs.

We took a random sample of examples of each type of scan and estimated the rate of OCR faults encountered; see Fig. 2. The results for the simulated scans hold intuitively, with ALLCAP and INITCAP scoring the lowest error rate, ALLCAP doing better due to larger clearer letters. then 1STLINE performs worst, nearly twice the rate of ALLCAP. But where as ALLCAP and INITCAP have similar query lengths, on average 1STLINE will be at least twice as long. Giving more room for errors. Here HAND scores highly even though it has an average query length similar to ALLCAP and INITCAP, which is consistant with above.

Fig. 3 shows actual examples of all five scan types.

3. EXPERIMENT 1: INDIVIDUAL SCANS

Our first experiment is designed to measure the effectiveness of each scan type in isolation. Fig. 4 shows, for each scan type, the percentage of catalogs for which the correct Web site is found at position x or better, as a function of x.

More precisely, let C be the set of all paper catalogs, let $c \in C$ be a specific catalog, and let site(c) be the authoritative Web site to which c corresponds. For some given scan type T, let $T(c)$ be the set of all scans of type T found in catalog c. For any $s \in T(c)$, let $r(s) = \{r_{s,1}, \ldots, r_{s,20}\}$ be the twenty results returned from Google, and let site(r_i) be the Web site of result r_i. Fig. 4 shows the *coverage* $Y(x, T)$, which measures the fraction of the time when we pick a random catalog and a random scan of type T that we can expect the correct result to be in position x or better:

$$Y(x,T) = \frac{1}{|C|} \sum_{c \in C} \left(\frac{1}{U} \sum_{1 \le u \le U,\ s \sim T(c)} [\![x \ge \min_{\mathrm{site}(r_{s,i}) = \mathrm{site}(c)} i]\!] \right),$$

where $[\![\rho]\!] = 1$ if ρ is true and 0 otherwise, $s \sim T(c)$ indicates that s is drawn randomly from $T(c)$, and we average over U random scans s in order to produce a stable estimate of $Y(x, T)$; in our experiments we use $U = 5$.

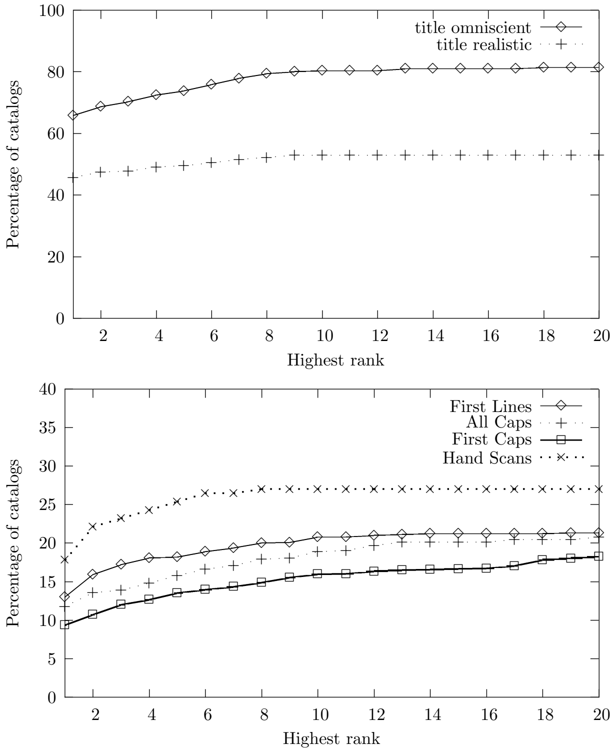

Figure 4: Coverage of TITLE (top) and other (bottom) scan types, as a function of the number of scans.

As expected, Fig. 4 shows that the TITLE scans do significantly better than the other types. The *title omniscient* curve is an upper bound on performance that uses the exact title for a given catalog regardless of whether it is present in the OCR output or not. The *title realistic* curve uses the correct title only if it is possible to retrieve it through OCR.

Note that HAND scans are better than automatically generated scans. This is to be expected: even though HAND scans have a more OCR errors per word, they tend to have a higher semantic relevance. This demonstrates that the results for our simulated scans serve as a lower bound on retrieval performance in a real application.

The experiments above treat the scanned text fragments as unquoted bags of words. We compared this technique with treating the fragments as quoted sequences of words. As shown in Fig. 5, the quoted scans produce the correct Web site at a better rank when it finds the right Web site, but also finds the right site less often. As intuition would suggest, quoting the scanned fragments increases their precision, but also increases the effect of OCR errors. The rest of our experiments will just treat scans as unquoted bags of words.

4. EXPERIMENT 2: MULTIPLE SCANS

The experiments discussed so far are all concerned with a single scan. If the user is willing to invest more effort by scanning additional text fragments, can the authoritative Web site be retrieved more efficiently?

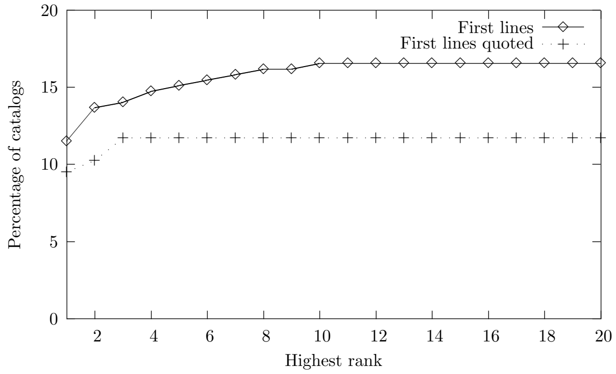

Figure 5: A comparison of quoted and unquoted search queries.

To explore this issue, we describe two approaches to combining the search engine results from multiple scans into a single ranked list of results. In each approach, we selected n scans randomly from either our three simulated scan types or from the HAND scans. Each scanned fragment s_i produces a list of 20 results $r_{s_i,1}, \ldots, r_{s_i,20}$, and we merge them into a new ranked list $R_1, \ldots, R_{20n}$. We report below the coverage versus search position, as above, but instead of reporting the entire graph, we plot n versus the search position between 1 and $20n$ at which we first achieve 50% coverage over the catalogs. We choose 50% as that is approximately the highest coverage achieved by the *title realistic* approach in Fig. 4.

Each additional scan phrase adds to the likelihood of finding the correct result, but also adds 20 positions onto the total range. So the challenge is to have each additional scan phrase add more to the overall accuracy than it looses due to the lengthening of the list.

4.1 Method 1: Naive merging

Our first approach merges the results from the scan phrases according to a simple weighting scheme. The simplest scheme would be to simply aggregate the lists so that there would be n ties for the first position, n ties for second, etc. We found that this simple scheme was ineffective, so we report results for the following weighted merging schema.

Consider a result $r_{s_i,j}$ from a scan s_i of type T. We generate the final ranked list by assigning a score to $r_{s_i,j}$ as follows: $\mathrm{score}(T,j) = 1 - (1 - \mathrm{pos}(j)) \cdot (1 - \mathrm{typerank}(T,j))$, where $0 < \mathrm{pos}(j) = (21 - i)/20 \leq 1$ is a score based on the original position j in the search results for scan type T. $\mathrm{typerank}(T,j)$ is score based on the chance that a scan of type T has the correct position at rank j or better: $\mathrm{typerank}(T,j) = \mathrm{prob}(T,j)/\sum_{T'} \mathrm{prob}(T',j)$, where $\mathrm{prob}(T,j)$ is the frequency (over a set of training data) that a scan phrase of type T at position j or higher is correct.

As shown in Fig. 6, this naive approach performs quite badly: to have a 50% chance of finding the authoritative Web site, more than 100 sample documents must be inspected before finding the correct one, even with many scanned

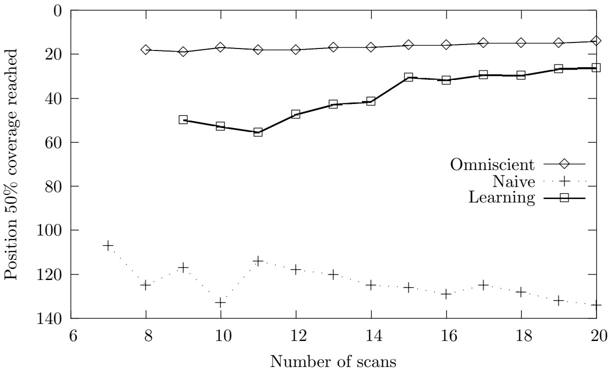

Figure 6: Comparison of the naive, learning and omniscient merging algorithms.

fragments. This is mainly due to the fact that a high proportion of the scanned fragments fail to find the correct site at any place in their 20 positions. This leads to numerous bad results being placed ahead of possibly good ones. To see this in more detail, consider the simple unweighted merging strategy. Each additional scan phrase s_i generates 20 results $\{r_{s_i,1}, \ldots, r_{s_i,20}\}$. If we have interleaved n scan phrases, we will now have a ranked list $\{R_1, \ldots, R_{20n}\}$ where each $r_{s_i,j}$ will be in positions $\{R_{1+n(j-1)}, \ldots, R_{nj}\}$ As can been seen, as we increase n we reduce the ranking of all results not originally ranked near the start, and since a large proportion of our correct results come from this region, the results degrade as new scans are added. In the case where the correct result is regularly near the start this will perform well. but in the case where the correct result is at a further position m, its new position will be in the vicinity of nm in the new ranked list, thus increases with n. So we need a way of promoting good results without merging parts of every scan phrase we add.

4.2 Method 2: Learning to identify good scans

While the previous approach does not perform well, there is good reason to be optimistic: intuitively, in a sufficiently large set of scanned fragments, there should be at least one "good" fragment (ie, a fragment whose query results yield the authoritative Web site near the start).

For example, consider an omniscient merging strategy that, given a group of scans, automatically "knows" which scan s_m has the authoritative Web site positioned earliest in its results. Then we can generate an upper bound on the coverage of the merged results by reserving positions $\{R_1, \ldots, R_{20}\}$ for s_m, instead of interleaving s_m with any other results. Fig. 6 shows that this omniscient approach performs remarkably well, and instead of suffering from our previous problem where additional scan phrases decrease coverage, there is a steady improvement as the number of scans increases.

Of course, the omniscient algorithm is not feasible in practice, but can we approach its performance? To answer this question, we adopted a machine learning approach in which we trained a classifier to recognise whether a set of results was likely to contain the correct result. To the extent that we can perform this classification task well, we can approach the performance of the omniscient algorithm. Given the problems with interleaved merging of results, here we treat the 20 results from each scan atomically: the results for scan s_i, $\{r_{s_i,1}, \ldots, r_{s_i,20}\}$, are added as a single block to the final merged list. We will use our classifier to calculate for each s_i the probability P_i that the scan s_i contains the site(c) at any position in its results. Our scans will then be sorted in order of these probabilities and concatenated to form the final merged list.

We anticipate that this approach will reduce the problem with the naive approach in which coverage decreased with additional scans. To illustrate, suppose we for have two scans, s_1 and s_2, where $P_1 = 0.6$ and $P_2 = 0.4$, which we merge into $R = r_{s_1,1}, \ldots, r_{s_1,20}, r_{s_2,1}, \ldots, r_{s_2,20}$. Then $P_{R_{40}} = 0.76$ where $P_{R_{40}}$ is the probability R contains site(c) within its first 40 positions. Lets assume s_a and s_b refer to the two currently highest ranked scans in R. Now suppose we add another m scans $s_3, \ldots, s_{m+2}$. For each scan s_i, $3 \leq i \leq m+2$, if $P_i > P_a$ then P_i will replace P_a and $P_{R_{40}}$ will as a result increase, otherwise if $P_i > P_b$ then P_i will replace P_b and again $P_{R_{40}}$ will increase. Otherwise s_i will be placed at some point further down in R having no effect on $P_{R_{40}}$. So by adding more scans $P_{R_{40}}$ may increase if a better scan is found, but it will never decrease. So unlike in the case of interleaved merging, the accuracy of our new ranked list is not adversely affected by the inclusion of 'bad' scans. The limitation is that as the top 20 elements of R always come from a single scan, we cannot improve the performance of the best single scan.

We used the C4.5 learning algorithm [9] to generate the classifiers. We identified a number of features for training.

scan_type: the type of scan feature which created this scan phrase;

num_under20: total number of results Google found (or 20, whichever is smaller);

num_total: total number of results Google found;

pairwise_sim $= \sum_{1\leq j<20} [\![\text{site}(r_{s_i,j}) = \text{site}(r_{s_i,j+1})]\!]$, where s_i is the scan whose results are being classified;

pairwise_sim2 $= \sum_{1\leq j\neq k\leq 20} [\![\text{site}(r_{s_i,j}) = \text{site}(r_{s_i,k})]\!]$;

freq_add: sum of the TF-IDF [1] frequencies of each individual term in our text fragment;

freq_prod: product of the TF-IDF frequencies;

freq_lowest: minimum of the TF-IDF frequencies; and

freq_highest: maximum of the TF-IDF frequencies.

The features **num_under20** and **num_total**, give a measure of how general versus how specific the results are. The

[1] TF-IDF is a standard information retrieval approach to assigning weights terms in text documents. The TF-IDF weight of a term t in a document d is the ratio of the frequency of t in d, dividing by the frequency of t in a large corpus of documents (eg, the entire Web) to which d belongs. The intent is that a document's high-weight terms will reflect its semantic content.

features **pairwise_sim** and **pairwise_sim2** give a measure of how likely the results have found one site especially more relevant. The TF-IDF (with the IDF value being generated from a large corpus of random Web documents) give a measure of whether the terms in the query are specialised or generic.

For feature selection, we simply generated every combination of features from the above features, learning classifiers for each, and picked the one which performed best. This happened to be the one using features **num_under20**, **pairwise_sim**, **pairwise_sim2**, **freq_add**, **freq_prod**, **freq_lowest** and **freq_highest**. We were pleased to discover that the classifier ignores **scan_type**; in a real-world application, this would suggest that the user does not need to associate a scan type with each fragment.

In our evaluation, we used three-fold cross-validation. That is, we randomly partition the catalogs into three sets S_1, S_2, and S_3, and then report the average accuracy over three separate learning tasks: training on $S_1 \cup S_2$ and testing on S_3, training on $S_1 \cup S_3$ and testing on S_2, and training on $S_2 \cup S_3$ and testing on S_1.

For each test set, we take n random scan phrases from the three simulated types and from HAND, and then calculate our scan phrase features. Using these and our classifier we assign a probability of 'goodness' to each scan phrase, and then merge them according to this probability. Due to the random nature of selecting scan phrases, we average over twenty repetitions of this procedure. The learned classifier has an accuracy of 84% (precision 68%, recall 41%).

Our results are shown in Fig. 6. We conclude that the learning approach is significantly better than the naive method. On average, after 20 scans, the correct site is found in the top 6% of the 400 results, compared to 34% for the naive method. Furthermore, we solved the problem of additional scans degrading performance.

4.3 Improving on Title

We investigated whether we could use these approaches to improve performance of the TITLE scans. To this end we ran the same ranked experiments from Sec. 3 with our random selection of scans always including one TITLE scan where available. Firstly we tried the interleaved ranking approach, and we suffered from the same problem as we encountered before. Namely with the majority of results in new scans being 'bad' by interleaved scanning we increasingly marginalise and good results not highly ranked to begin with.

Since we already have a good result first (the TITLE) we will get less noticeable gain from merging in an atomic manner, But the inclusion of one scan of higher quality serves as a boost to all the results. Our results here are still provisional, so we have not included them with this paper.

5. RELATED WORK

This task of mapping scanned text fragments to authoritative Web sites is a mixture between a database merging problem and a search problem. We have separate ranked lists for each individual scan, and the goal is to merge them into a single ranked list. There are many approaches to aggregating ranked lists (eg, [4, 3] These solutions mainly involve merging ranked lists from independent sources to ensure adequate performance. Since our problem involves merging ranked lists from highly dependant sources, and we are not looking for a better ranking of results of a broad type but for improving the ranking of a specific Web site, these strategies do not map well to our task.

This case study also involves issues of document analysis / recognition. There is a large body of work concerning OCR and scanning technology in general (eg, [2] or the annual Electronic Imaging conferences[6]), but we know of no prior work that focuses on this kind of task.

The challenge of linking paper and digital media has received substantial attention in the ubiquitous computing and human computer interaction literature (eg, [5, 1, 7, 8, 11, 10]), but none of this work appears to be directly relevant to our task. For instance, in the efforts to link physical and Web artifacts [5, 1, 7], it is assumed that an explicit link between a physical object and its digital equivalent is stored in, for example, digitally enhanced 'paper'. While such enhancements may be realized in some situations, part of the reason paper is still so commonly employed is due to its cheapness and resilience, so there will continue to be a need to map ordinary paper artifacts to their Web counterpars. To the best of our knowledge, no one is currently working on ranked merging of results to locate a specific Web site.

6. DISCUSSION

Paper has been used extensively for centuries, so it is hardly surprising that the recent and growing surge in digital media has supplemented paper rather than replaced it. The fact that we have the option of either a paper or Web instantiation of most documents is beneficial from the user's perspective: they each have distinct advantages, and people have their own preferences. Motivated by the vision of seamless integration across the two media, our goal is to develop technology so that the two media can be used interchangeably. Unfortunately, current knowledge management, ubiquitous computing and digital library technologies do not support this vision adequately.

An important enabling technology is the ability to locate the authoritative Web document that is equivalent to a given physical artifact. To investigate this issue, we performed a case study on a sample of 295 catalogs electronically scanned from their original paper catalogs. We simulated a variety of scenarios in which a user has a physical paper catalog and a pen scanning device, and uses scanned fragments of text as queries to a Web search engine in order to retrieve the catalog's authoritative Web site. Since our goal is to minimise the effort required to perform this task, we measured the retrieval effectiveness as a function of the number of scanned text fragments, and we investigated several techniques for combining these.

We are currently improving the effectiveness of our techniques in several ways. One obvious possibility for reducing the required number of scans by reintroducing both the quoted and unquoted versions. Another possibility would be to identify, for our classifier, features across separate scan phrases from the same catalog. We also intend to address the more challenging task of finding the exact resource (rather than just the root) of the authoritative Web site to which the paper catalog corresponds.

Ultimately, our goal is to develop techniques so that the paper catalog can be used as a medium for interacting with the actual Web service. For example, we are developing a tool that will enable a person with access to an on-line e-commerce Website, to automatically select and add items to their shopping cart, using the PDA and pen scanner as an input device. To this end we are attempting to design a system that can use an generic model of an e-commerce website to learn how to interact with an unseen specific e-commerce website. We are also considering applications related to organisational workflow processes, such as filling out an expense claim form based on a data from a combination of paper and Web sources.

Acknowledgements. This research was supported by grants SFI/01/F.1/C015 from Science Foundation Ireland, and N00014-03-1-0274 from the US Office of Naval Research. We thank Google for their product catalog service.

7. REFERENCES

[1] T. Arai, D. Aust, and S. Hudson. Paperlink: A technique for hyperlinking from real paper to electronic content. In *Proceedings of ACM CHI'97*, 1997.

[2] A. Chhabra, L. Schomaker, and J. Kim, editors. *7th International Conference on Document Analysis and Recognition (ICDAR 2003), Edinburgh, Scotland, UK.* IEEE Computer Society, 2003.

[3] William W. Cohen, Robert E. Schapire, and Yoram Singer. Learning to order things. In *Advances in Neural Information Processing Systems*, volume 10. The MIT Press, 1998.

[4] Nick Craswell, David Hawking, and Paul Thistlewaite. Merging results from isolated search engines. In *Australasian Database Conference*, 1999.

[5] Kaj Gronbaek, Jannie F. Kristensen, Peter Orbaek, and Mette Agger Eriksen. Physical hypermedia: Organising collections of mixed physical and digital material. In *Proceedings of Hypertext'03*, 2003.

[6] IS&T/SPIE. *IS&T/SPIE's 15th Annual Symposium on Electronic Imaging*, 2003.

[7] W. Johnson, H. Jellinek, L. Klotz, R. Rao, and S. Card. Bridging the paper and electronic worlds: The paper user interface. In *Proceedings of INTERCHI'93*, 1997.

[8] K. O'Hara and A. Sellen. A comparison of reading paper and on-line documents. In *Proceedings of CHI-97, Special Interest Group on Computer & Human Interaction.*, 1997.

[9] Ross J. Quinlan. *C4.5: Programs for Machine Learning.* Machine Learning. Morgan Kaufmann, 1993.

[10] Arai T., Machii K., and Kuzunuki S. Retrieving electronic documents with real-world objects on interactivedesk. In *Proceedings of UIST'95*, pages 37–38, 1995.

[11] Pierre Wellner. Interacting with paper on the DigitalDesk. In *Communications of the ACM*, pages 86–97, 1993.