AllInOneNews: Development and Evaluation of a Large-Scale News Metasearch Engine

King-Lup Liu[1], Weiyi Meng[1], Jing Qiu[2], Clement Yu[1], Vijay Raghavan[1], Zonghuan Wu[1]
Yiyao Lu[1], Hai He[1], Hongkun Zhao[1]
[1]Webscalers, LLC, Lafayette, LA 70506, USA, {kliu, meng, yu, raghavan, wu}@webscalers.com
[2]SUNY at Binghamton, Binghamton, NY 13902, USA, jqiu2@binghamton.edu

ABSTRACT

AllInOneNews is the largest news metasearch engine in the world, connecting to over 1,000 news sites over 150 countries. Implementing a large-scale metasearch engine like AllInOneNews needs to overcome unique challenges not faced by building small metasearch engines such as developing highly scalable search engine selection techniques. In this paper, we discuss these unique challenges and our solutions to these challenges. We also discuss some novel features of AllInOneNews such as highly automated solution and semantic query match. This paper also reports the results of a comparative evaluation of three commercial news search systems, one search engine – Google News and two metasearch engines – Mamma News and AllInOneNews. Several measures such as effectiveness, diversity and time-sensitivity are used to perform the comparison. Another contribution of this paper is that we introduce a novel scheme to compare multiple news search systems in a combined measure that takes both relevance and time-sensitivity of retrieved information into consideration.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *retrieval model*, *search process*, *selection process*; H.3.4 [**Information Storage and Retrieval**]: System and Software – *Performance evaluation (efficiency and effectiveness)*; H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *Web-based services*.

General Terms

Design, Measurement, Performance, Algorithms

Keywords

Metasearch engine, news search, search engine, time-sensitive ranking

1. INTRODUCTION

Search engines and metasearch engines are two different types of search systems on the Web. The former crawls documents from the Web and builds a local index to support user search and the latter passes user queries to other search engines, collects the search results from these search engines and merges them into a single ranked list for presentation to their own users.

While search engines like Google and Yahoo have achieved great success, metasearch engines have some unique advantages over search engines such as their abilities to reach the deep Web (by connecting to the search interfaces of deep web sources), to combine the coverage of multiple search engines and to produce more accurate results (a good result merging algorithm can produce better results because it can pick the overall best results among the best results from individual search engines [14]). Technical issues for building metasearch engines have been widely studied (e.g., [1, 4, 6, 7, 8, 9, 10, 13, 16, 17, 21, 22, 23, 26] (Please see [19] for a survey on metasearch engine technologies) and many commercial metasearch engines have been built (e.g., Dogpile (www.dogpile.com), Vivisimo (www.vivisimo.com) and Mamma (www.mamma.com)). Most of the current best-known metasearch engines connect to only a dozen or so general-purpose search engines.

This paper focuses on news search systems. Many kinds of information are sensitive to time. This is especially the case with news. The value of a news article depends much on its time of publication. Today, an increasing number of people are reading news online. Online news is mostly free and easily accessible with a web browser. A more important reason for its popularity is that people can obtain recent news as well as old news that may not be readily available from newspapers. News articles can be accessed as soon as they are posted. Also, with the search capability of a news web site, people can readily locate news items that are of interest to them. However, there are many news organizations in the world and there are also many specialized news sites, such as those sports, finance, entertainment, and local (ones that cater to local communities). A news article of interest to a person may be posted in a newspaper web site unknown to him/her. To solve this problem, a number of news search engines, notably Google News (news.google.com), have been created that allow people to search news articles from a number of news organizations around the world from a single search system. These news search engines were constructed using conventional techniques. That is, they periodically send "web crawlers" to fetch news articles from the news organizations, parse them and then update the document indexes at their servers. A news article is treated as if it were a regular web page posted on the Internet.

AllInOneNews (www.allinonenews.com) is a news search system based on the metasearch engine technology. In other words, AllInOneNews is a news metasearch engine. AllInOneNews connects directly to the search engines of individual newspapers and news sites. Like Google News, AllInOneNews also allows its users to search news from multiple news sources. However, we believe the AllInOneNews technology may yield the following potential advantages over a crawler-based news search engine: (1) it can return more recent news because as soon as a news article becomes available at a local news site, it can be searched by AllInOneNews while the crawling delay may make it not immediately accessible by a crawler-based search engine; (2) it can reach old news articles by connecting to the search interfaces of archive news search engines; in contrast, crawling based

SIGMOD'07, June 11–14, 2007, Beijing, China.
ACM 978-1-59593-686-8/07/0006.

technique cannot crawl news articles available only via search interfaces (Google's new News Archive Search uses non-crawling techniques to obtain old news articles such as getting them directly from newspaper publishers through business relationships); and (3) it can be more scalable in terms of covering more newspapers/news sites as there are much fewer news search engines than the number of news items.

This paper has three main objectives. The first is to introduce the novel features of AllInOneNews and to provide some insights on how AllInOneNews is developed. The second is to propose a set of criteria for evaluating the quality of news search systems. The third is to report the results of our comparative evaluation of three news search systems, namely Google News (news.google.com), Mamma News (www.mamma.com/MammaNews), and AllInOneNews (www.allinonenews.com). Google News is a search engine that has a centralized index (which could be replicated at multiple locations). Both Mamma News and AllInOneNews are metasearch engines.

In conventional information retrieval, the retrieval effectiveness of a system is usually measured by two quantities, *recall* and *precision*, where *recall* is the percentage of the relevant documents that are retrieved and *precision* is the percentage of the retrieved documents that are relevant. With these two measures, all relevant documents are treated as having the same importance. But for time-sensitive information, this should not be the case. Consider news report. In many cases, a relevant news article is considered to be more important that an older relevant one. For a disaster such as a tsunami in progress, an up-to-date report is much preferred than another one that is half an hour late. Thus, for time-sensitive information, the importance of a relevant document should also be judged based on its time of publication. Hence, some new measures need to be used to evaluate the effectiveness of systems that retrieve time-sensitive information. In this paper, we introduce one such measure. In the literature, we are not aware of any reports of similar studies. We use this new measure of relevance of information to compare the retrieval effectiveness of systems that retrieve time-sensitive information: the three news search/metasearch engines mentioned above.

The contributions of this paper are:

1. We provide some insights on implementing a large-scale news metasearch engine, AllInOneNews, and introduce many of its novel features. No such insights have been provided on any other large-scale metasearch engines before.
2. We provide several evaluation criteria to measure the quality of news search systems. In particular, we propose a new and novel scheme to measure and compare the retrieval effectiveness of multiple search/metasearch engines for time-sensitive information.
3. We compare three search/metasearch engines using the developed criteria. Furthermore, we identify the strengths and weaknesses of these systems in processing different types of queries.

The rest of this paper is organized as follows. In section 2, we give a brief description of related work. Section 3 introduces the three news search systems we plan to compare, with the emphasis on AllInOneNews. In section 4, we propose the criteria for evaluating news search systems. In section 5, a description of our experiment is given, and we also discuss the strengths and weaknesses found in the three news search/metasearch engines. We summarize and conclude this paper in section 6.

2. RELATED WORKS

While many metasearch engines have been developed and deployed on the Web and some of them also support news search, there is little published information on their implementation. For those there is published information (such as ProFusion and SavvySearch), the information is somewhat old and is not related to news search. Furthermore, most publicly accessible metasearch engines are small with only a dozen or so search engines. Some metasearch engines have more search engines but these search engines are grouped into different categories in advance and for each category, only a small number of search engines are used. In a sense such a metasearch engine can be considered as a combination of multiple small metasearch engines. In contrast, AllInOneNews is a large-scale metasearch engine connecting to over 1,000 search engines. In addition, AllInOneNews is built using highly automated solutions for many of its components. We are not aware of any other public metasearch engines that have similar scale and are constructed based on mostly automated solutions. AllInOneNews also has other unique features not reported in other metasearch systems such as semantic-based query match and time-sensitive result merging.

Whenever a new retrieval technique is introduced, people typically compare the retrieval effectiveness of a system designed using the new technique against some other information retrieval systems. As the information stored in many information retrieval systems are mostly not time-sensitive, the comparison seldom takes into account the recency of the content of the retrieved documents. In [20], a comparison of several result merging strategies was conducted for a news metasearch engine. This can be viewed as a comparison of different news metasearch engines each using a different result merging strategy. In the study in [20], the news articles returned were not obtained directly from the Web. Instead, they were from a database of news articles obtained by submitting 114 queries to 15 news organizations and fetching the top 10 news articles in the result list returned for each query. The comparison performed was still using conventional information retrieval technique and did not consider the recency of each returned news article as well as other quality measures. To our knowledge, our work is the first on the evaluation of information retrieval systems that also take into account the recency of the content of retrieved information.

3. THREE NEWS SEARCH SYSTEMS

Since both Mamma News and AllInOneNews are metasearch engines, we first provide a brief overview of the general metasearch technology before introducing the three news search systems.

3.1 Metasearch Technology

A metasearch engine is a system that provides unified access to multiple existing search engines. From a user's perspective, there is essentially no difference between using a search engine and using a metasearch engine. However, search engines and metasearch engines are built using very different techniques [25].

When a metasearch engine receives a query from a user, it sends the query to multiple existing search engines (which will be called *component search engines* in this paper), and it then combines the results returned by these search engines and displays the combined results to the user. A metasearch engine makes it easy for a user to search multiple search engines simultaneously while submitting just one query.

A simple metasearch engine consists of a *user interface* for users to submit queries, a *search engine connection component* for submitting queries to its component search engines and receiving result pages from them through programs, a *result extraction component* for extracting the search result records (SRRs) from the returned result pages, and a *result merging component* for combining the results [19]. If a metasearch engine employs a large number of search engines, then a *search engine selection component* is needed. This component determines which search engines are likely to contain good matching results for any given user query so that only these search engines are used for this query. Search engine selection is important mostly for efficiency. For example, suppose only the 50 best-matched results are needed for a query and a metasearch engine has 1,000 component search engines. Clearly the 50 best-matched results will come from at most 50 component search engines, meaning that at least 950 of the 1,000 search engines are not useful for this query. Sending a query to useless search engines will cause serious inefficiencies, such as heavy network traffic caused by transmitting unwanted results and the waste of system resources for evaluating the query. Furthermore, the metasearch engine may be overwhelmed by the irrelevant results returned by the useless search engines.

For a news metasearch engine, a new component – the *publication time extraction component* – is needed to identify the publication date and time of each retrieved news item. Publication time is needed to perform time-sensitive ranking of retrieved results. In general, among the relevant news articles, more recent ones should be ranked ahead of older ones.

In section 3.2, we will discuss how the above main components are implemented for AllInOneNews.

3.2 AllInOneNews

AllInOneNews is implemented based on the metasearch engine technology. Currently it is connected to over 1,000 English news search engines from about 150 countries. These news search engines are mostly selected from several newspaper lists on the Web such as http://www.newspaperindex.com, http://www.world-newspapers.com and http://www.onlinenewspapers.com. Only sites that satisfy certain requirements are used. For example, sites that do not have search engines are generally not used (we created search engines for a small number of such sites to allow them to be used by AllInOneNews). As another example, if the average response time of a site over several sample queries exceeds a pre-set threshold, it will not be used.

The final goal of AllInOneNews is to connect to all news search engines in multiple languages on the Web. At the same time, we aim to make AllInOneNews the most effective news search system on the Web. To accommodate the large number of component search engines, highly scalable and automated solutions are developed to implement and maintain AllInOneNews. To support effective retrieval, several advanced features are incorporated into AllInOneNews. Below, we provide a brief summary about the implementation of the main components of AllInOneNews and provide insights on how good scalability and effectiveness are achieved. We should point out that while AllInOneNews is developed based on the principles and main ideas described below, the actual implementation involves additional innovations and details.

3.2.1 Automatic Search Engine Connection

Usually, the search interface of a search engine is implemented using an HTML *form tag* with a search textbox. The form tag contains all necessary information needed to connect to the search engine through a program. Such information includes the name and the location of the search engine server that processes user queries, the network connection method (i.e., the HTTP request method, usually GET or POST), and the internal name associated with the search textbox (the query string is assigned to this name when the query is passed to the search engine server). A form parser can be used to automatically extract the above information from the HTML interface page of a given search engine. Once the above basic pieces of information are obtained, it is not difficult to generate a connection program for the search engine for passing queries and receiving results.

While a basic form parser can be implemented reasonably easily, developing a truly robust form parser is quite challenging due to many complexities involved such as differentiating search forms from non-search forms, dealing with multiple search forms on the same interface page, dealing with redirection and handling Javascript. Let's discuss one issue in some detail. Consider the issue of differentiating search forms from non-search forms. This problem occurs because a given interface page may have more than one form tag and not every form tag corresponds to a search engine. For example, login/sign-in interfaces and many online surveys are also HTML forms. In [22], the following solution is provided to identify search engine forms: first the form must have a textbox; second, one of a set of pre-compiled keywords such as "search", "find" and "seek" must appear either within or near the form tag; third, generate a connection program based on the extracted connection parameters and send test queries to see if any results can be returned, if yes, then the search form is confirmed. An extended version of this solution is used in the implementation of AllInOneNews. Machine learning algorithms for detecting search forms can be found in [3, 5].

At present, AllInOneNews' automatic search engine connection component cannot handle search forms with heavy Javascript. Consequently, news search engines whose search forms have heavy Javascript are not used.

3.2.2 Search Engine Selection

For a metasearch engine with as many component search engines as AllInOneNews has, having an effective search engine selection algorithm is imperative. When a user query is received by AllInOneNews, its search engine selection algorithm identifies several dozens of component search engines that it believes are most suitable for this query and only these search engines will be used to evaluate this query. Below, we outline the main ideas behind AllInOneNews' search engine selection component.

Before any meaningful search engine selection can be performed, some information representing the contents of the set of pages of each search engine needs to be collected and this information is called the *representative* of the search engine [19]. The representatives of all search engines used by the metasearch engine are collected in advance and are stored with the metasearch engine. During search engine selection for a given query, search engines are ranked based on how well their representatives match with the query. Many search engine selection (also called database selection) algorithms have been proposed in the literature and they differ by the kind of representatives they use and the ways in which the representatives are used [19].

The search engine selection algorithm adopted by AllInOneNews is a revised version of the optimal ranking algorithm described in [18, 24]. This method is summarized below:

1. First, a representative is obtained for each component search engine S. For each component news website, we regularly crawl for new news articles and use these articles to compute its representative. For each distinct term t in S (i.e., t appears in an article in S), a statistical information called the *adjusted maximum normalized weight* amw(t, S) is computed for t as follows: compute the normalized weight of t in every document (i.e., the term frequency weight of t divided by the length of the document) in S, find the maximum value among these weights, and multiply this maximum weight by the global *idf* weight of t across all component search engines. Let G be the number of news articles across all component news search engines that contain t, i.e., G is the global document frequency of t, then the global *idf* weight of t is computed by log(N/G), where N is the total number of news articles across all search engines.
2. Second, the representatives for all component search engines are integrated into a single representative. For each term t, among {amw(t, S_i): i=1,2, …}, the r largest values are kept and the rest are discarded for a small integer r (say $\leq$ 50). The idea is that if a search engine does not have one of the r largest weights for t, then it is unlikely to be among the most useful search engines for a query containing t. If amw(t, S_k) is kept for t, then the id# of search engine S_k is kept as well. In other words, up to r pairs (amw(t, S_k), id#(S_k)) are kept for t in the integrated representative.
3. Third, with the objective of retrieving the m most similar (relevant) documents with respect to a given query Q from across all component search engines, we say the search engines are optimally ranked with order [S_1, S_2, …, S_n] if for any m we can find a k such that the m most similar documents are contained in [S_1, …, S_k] and each of these k search engines contain at least one of the m most similar documents. A necessary and sufficient condition for the component search engines to be optimally ranked is to order the search engines in descending order of the similarity of the most similar document with respect to Q in each search engine.
4. Fourth, while techniques exist for estimating the similarity of the most similar document in any search engine S with any given query Q using the information in the database representative, they are not very efficient. Instead, we aim to efficiently estimate a quantity for any S and Q such that when the search engines are ranked in descending order of this quantity, their order will be close to the optimal order. One such quantity is max{amw(t, S): $t \in$ Q} [18]. Based on this, the following efficient search engine ranking process can be employed: for query Q, first identify the r pairs of {amw(t, S_k), id#(S_k)} for each term t in Q; then we compute the above quantity for each search engine whose id# is in the above pairs; finally rank the search engines in descending order of these quantities. If Q has w distinct terms, then at most $r*w$ quantities need to be computed, independent of the number of component search engines. For a typical Internet query with 2-3 terms and for r = 30, only 60-90 simple computations are needed to rank the component search engines approximately optimally.

Note that even though the integrated representative used by AllInOneNews is generated through crawled news articles, the metasearch approach is still different from crawling based search engines because the representative is far less sensitive to changes than the documents themselves. For example, once a term t about an event has been added to the integrated representative with respect to a site, say based on an earlier news article from the site, t can be used to retrieve fresher news articles about the event from the site even when these fresher articles have not been crawled.

3.2.3 Automatic Wrapper Generation

Result pages returned by a search engine are dynamically generated HTML pages. In addition to the search result records (SRRs) for a query, such a result page usually also contains some unwanted information/links such as advertisements, search engine host information and sponsored links. It is essential to correctly extract the SRRs on each result page. A typical SRR corresponds to a Web page found by the search engine and it usually contains the URL, the title and a snippet of the page.

A wrapper in this paper refers to the rules that can be used to extract the SRRs from the result pages returned from a search engine. The wrapper for different search engine is usually different, which means we need to generate a wrapper for every component search engine. While it is possible to generate wrappers manually or semi-automatically, doing so for a large number of component search engines can incur very high labor cost. AllInOneNews employs an automatic wrapper generation tool (ViNTs [27]) to generate the wrappers. For search engines whose wrappers cannot be automatically and correctly generated, a semi-automatic wrapper generation tool is utilized. Below, we provide a brief overview of the ViNTs system.

ViNTs takes one or more sample result pages from a search engine as input, which can be provided either manually by a user or automatically by the system through submitting automatically generated sample queries to the search engine. For each input sample result page, its DOM tree is built to analyzing its tag structures and it is rendered on a browser to extract its visual information. *Content-line* is the basic building block of the ViNTs approach. Specifically, a content line is a group of characters that visually form a horizontal line in the same section on the rendered page. In ViNTs, eight types of content lines (such as link line, text line, blank line, etc.) are differentiated. Next, the content lines are divided into *blocks* based on candidate *separators* which are repeating content lines. Candidate separators that yield similar blocks are kept. The similarity between two blocks B1 and B2 is defined in terms of their *type distance* between the type sequences of the content lines of B1 and B2, *shape distance* between the left contours formed by the left most x coordinates of the content lines in B1 and B2, and *position distance* between the left most x coordinates of the two blocks. Next, candidate wrappers (regular expressions of tags) are generated based on consecutive candidate blocks.

Note that a result page may have multiple sections and more than one section may contain neatly arranged records. For each candidate wrapper, ViNTs next determines the boundaries of the records it should extract. This is called *wrapper refining*. If more than one section has neatly arranged records, then more than one candidate wrapper may be generated after wrapper refining. Next ViNTs selects one of the candidate wrappers as the wrapper for this Web page (at present, ViNTs assumes each result page has only one main result record section, which is true for most search engines). ViNTs uses several heuristics to select the final wrapper for the page, including the size of the section as determined by its records (the main section usually has the largest area compared to other sections), the location of the section (the main section is usually centrally located on the result page), etc.

Many search engines do not display their search results in the same format. For example, Google displays some records un-indented while some indented. To increase the likelihood of capturing all varieties, multiple sample result pages from the same search engine should be used to generate the wrapper. For each sample page, the process described above is followed to generate a wrapper for the page. Then the wrappers for different sample pages are integrated into a single and more robust wrapper for the search engine. In ViNTs, wrapper integration includes *tag path integration* (which is the tag path from the root of the DOM tree to the root of the lowest subtree that contains all the search records) and *separator integration*.

ViNTs is specifically developed for search engines and it is also the first fully automatic wrapper generation system that utilizes both visual contents and tag information for wrapper generation. We believe it is one of the most accurate automatic wrapper generators currently available (its record level accuracy is about 98%). If an acceptable wrapper cannot be generated for a news search engine, the search engine will not used in AllInOneNews.

3.2.4 Publication Time Extraction

News is among the most time-sensitive information because more recent news on a topic is usually considered to be more useful. For AllInOneNews to be effective, it is critical that among all relevant/similar news articles for a given query, the more recent ones are ranked ahead. Therefore, how to extract the publication time of news result records returned from different news sources is a critical task. Here we describe how publication time is extracted from the search result records (SRRs) in AllInOneNews.

We first explain the issues associated with the publication time extraction problem. The first is the heterogeneities among the news sources in displaying the publication time. Different news sites may choose different locations in the SRR to display/encode the publication time information. Some prefer putting it close together with the article title while others like to put it in the snippet or even in the URL. The second is that publication times generated by different news sources may consist of different time components. A complete time may consist of year, month, day in a month, hour, minute, second, AM/PM, and even the time zone, but different news sources often adopt different subsets of these components in their publication times. The third is that different news sites may interpret the same time expression differently. The ambiguity might be caused by the difference in conventions of different countries. For example, the expression “04/03/07” could be interpreted as “April 3rd, 2007”, “March 7th, 2004” or “March 4th, 2007”, largely depending on from which country the news originally comes from. Finally, an SRR may contain multiple times and not all of them are the publication time.

AllInOneNews’ publication time extraction component addresses all of the above issues and it can be summarized below (the detail is reported in [15]):

1. First, time and date from each SRR, if exists, are extracted based on a set of pre-compiled time/date patterns, including special patterns that only appear in news SRRs such as “xx minutes ago” and “…/dd/dd/dddd/…” encoded in some URLs. For each date/time extracted, its location in the SRR is also recorded. For example, it can be at the beginning/middle/end of the title/snippet or in the URL.
2. Next, time and date values are “normalized” to a standard form using a variety of techniques. For example, time/date conventions in different countries and time zone information are utilized to help the normalization; “xx minutes ago” is converted to a date and time by subtracting “xx” minutes from the current date and time when the SRR is retrieved; “November 16” will be interpreted as “November 16, 2006” if the current date is on November 16, 2006 or after, otherwise it will be interpreted as “November 16, 2005”. In addition, for an expression like “04/03/07”, if another expression from the same site has “19/11/06” which implies the first two digitals must represent day not month (month can only be between 1 and 12), we can interpret “04/03/07” as March 4th, 2007.
3. Identify the publication time (including date) from the extracted times for each SRR. First, very old times and times for the future are removed because they cannot be publication times. Second, group all the remaining times from all SRRs from the same search engine based on their locations, e.g., all times appearing at the beginning of the title will be placed into the same group. The times in the largest group are recognized as the publication times based on the observation that news sites usually place the publication times at the same place in all of its SRRs.
4. For each news search engine, a publication time extraction wrapper is generated automatically based on sample SRRs extracted from the result pages returned from this search engine. The wrapper contains pattern, location and interpretation information for the publication times returned by this search engine. Not only the wrapper can be efficiently applied to extracting the publication times from newly returned SRRs from the same search engine (the above complicated process can be avoided), it also leads to better accuracy.

Based on our experiments, the algorithm described above has close to 100% accuracy.

3.2.5 Result Merging

The quality of the result-merging algorithm employed by a metasearch engine probably has the most direct impact on the effectiveness of the metasearch engine. Many result-merging algorithms have been proposed in the literature (e.g. [14]) and some are proposed specifically for the news metasearch engine context [20]. Due to the fast response required of metasearch engines, most result-merging algorithms do not download the actual web pages retrieved and compute their similarities with the query. Instead, they use only information available on the already returned result pages, including the rank of each result record from the search engine that returned it, the title and snippet contained in the result record. The similarities of the title and snippet of a result record with the query can be computed and aggregated to produce an overall ranking score for the record. In addition, the proximity of the query terms in the title and snippet can also be taken into consideration [14].

AllInOneNews employs a sophisticated result merging technique. This technique takes many factors into consideration when determining the rank of each result. Some of these factors are: the quality of selected component news search engine (which was obtained in the search engine selection step) from which the result is retrieved, the number of terms that are common between the query and the title/snippet of the result, the proximity of the query terms in the title/snippet of the result, the order in which the query terms appear in the result, and the publication time of the result.

3.2.6 Other Significant Features

AllInOneNews puts a lot of emphasis on recognizing and utilizing phrases as they often convey more precise information than their component words individually. Phrases are recognized from the sample news articles used to generate the representatives for the search engines based on proximity of words and the correlation of words (words in phrases usually appear close to each other and are highly correlated). These phrases are considered as special terms in the representatives and they are utilized in result merging. They are also utilized to recognize phrases in user queries so phrase-based match can be conducted.

Another significant feature of AllInOneNews is its support of semantic-based matching. For example, if a user query is “email”, results that contain “e-mail” and “electronic mail” can also be retrieved because they all have the same meaning. As another example, for query “renal calculus”, documents with “kidney stones” will also be matched. AllInOneNews uses a combination of several techniques to identify terms that are semantically similar to query terms, including utilizing semantic dictionaries like WordNet.

3.3 Mamma News

Mamma is a publicly traded company that specializes metasearch engines. Mamma News is part of the Mamma metasearch engine and it is also implemented based on the metasearch engine technology. It is not clear how many news search engines it employs. There is little publicly available information about the techniques (such as the result-merging algorithm) behind Mamma News.

3.4 Google News

Google News is part of Google and it is a news search engine implemented using a crawler-based technology in that it crawls news items from various news sites and builds a central index to answer users’ queries. By the time this paper is being written, Google News crawls news items from 4,500 English news sources. While it is known that Google News ranks news items based on the information in the title and text as well as the publication time, the details of the ranking algorithm are not known to the public.

4. QUALITY MEASURES

In this section we propose five criteria that we believe are important to measure the quality of news search systems. In Section 5, we will use these criteria to compare the three news search systems described in Section 3.

4.1 Traditional Effectiveness

Recall and precision are the most widely used quantities for measuring information retrieval systems. However, it is difficult to compute recall when evaluating search engines because it is difficult (impossible in many cases) to know the number of relevant documents for queries. One popular measure for evaluating the effectiveness of search engines is the TREC-style average precision (TSAP) [12]. TSAP at cutoff N (only the top N results are considered), denoted as TSAP@N, is defined below:

$$TSAP@N = \left(\sum_{i=1}^{N} r_i\right) / N$$

where $r_i = 1/i$ if the i-th ranked result is relevant and $r_i = 0$ if it is not relevant. This measure takes both the number of relevant documents and their positions in the result list into consideration.

For example, suppose for a given search engine and a query the relevance information of the top 5 retrieved documents is (R, I, I, R, R), where an R (I) at the i-th position indicates that the i-th result is relevant (irrelevant, respectively). Then for this query, the TSAP@5 value for this search engine is:

$$TSAP@5 = \frac{1/1 + 0 + 0 + 1/4 + 1/5}{5} = 0.29$$

4.2 Time-Sensitive Effectiveness

One thing unique about a news item is that its value often depends on its time of publication. Therefore, it is important to evaluate news search systems based on how well they retrieve fresh news items. We did not come up with a good measure that combines both time-independent effectiveness and the relative freshness of news items. Instead, we devised a scheme that can be used to compare the time-sensitive effectiveness of multiple news systems. The basic idea is to put all the relevant results from the news search systems we want to compare together into a single ranked list based on their recency and analyze the relative positions of these results. The details of this scheme are described below.

We assume that news items are time-stamped, and that the relevant news items can be arranged in non-ascending order of their timestamps. We also assume that the importance of a news item can be given by the Zipfian distribution. Specifically, if a relevant item has rank r, its degree of importance is given by $1/r^p$, where $0 \le p \le 1$. Note that when $p = 0$, each relevant item has the same importance. The value of p should be adjusted based on the sensitivity of the news items with respect to time (larger value indicates higher sensitivity). The rank of a relevant news item is j, if it is the j-th most recent news item among all the news items that are relevant to the query. It assumes that the more recent a news item is the more important it is. If the user wants a different ranking, the methodology given below is also directly applicable. The experimental result reported in this paper assumes that the news items are ranked in descending order of time. In other words, more recent ones have higher ranks. If an item is irrelevant, its degree of importance is 0. For all the relevant news items retrieved from all the news search/metasearch engines that are to be compared, we compute the sum of their degrees of importance. Let this sum be REL_TOTAL. Consider the top n news items in the result lists returned for a query from a news search/metasearch engine. We define the recall of these n news items to be REL_SUM(n) / REL_TOTAL and the precision to be REL_SUM(n) / n, where REL_SUM(n) is the sum of the degrees of importance of the top n news items.

Suppose that a query Q has been sent to three search/metasearch engines S1, S2 and S3. Suppose further that the result list returned from S1 is (I, R1, R4), the result list returned from S2 is (R2, I, R4), and the result list returned from S3 is (I, R3, I), where Rj is a relevant news item having rank j among all relevant news items retrieved from the three systems, and the I’s are the irrelevant news items. For example, for the result list returned from S1, the first item is irrelevant; the second item is relevant and is the most recent item among all relevant items from all of the three search systems; and finally the third item from S1 is also relevant and is the fourth most recent item among all relevant ones. In this example, the same document with rank 4 is returned from both S1 and S2. Then, the sum of the degrees of importance of all the relevant news items is:

$$\text{REL\_TOTAL} = \frac{1}{1^p} + \frac{1}{2^p} + \frac{1}{3^p} + \frac{1}{4^p}.$$

The recall of S1 for the first two news items, for example, is $(1/1^p)/REL\_TOTAL$ and the precision is $(1/1^p)/2$ where $1/1^p$ is due to R1. If the first three news items from S1 are considered, the recall of S1 is $(1/1^p + 1/4^p)/REL\_TOTAL$ and the precision is $(1/1^p + 1/4^p)/3$.

After each news search/metasearch engine retrieved the same number of documents, the total number of relevant documents is assumed to be the number of distinct relevant documents in the sets of retrieved documents.

4.3 Redundancy

Due to the sharing of news stories among many news sources, the same news item may appear in multiple newspapers. Consequently, for news search systems that gather news articles from multiple news sources, like the three news search/metasearch engines to be compared in this paper, their results may contain duplicate news items. A good news search system should make an effort to reduce or eliminate the redundancy. In this paper, two news items are considered to be the same if they have the same contents, including pictures and hyperlinks. For a given result set R, its degree of redundancy is defined as below:

$$\text{Redundancy} = 1 - K / |R|,$$

where K is the number of unique results in R and |R| is the size of R. Search engines with higher average scores produce more redundant results.

4.4 Diversity

It is understandable that different news articles about the same event can be very different because different reporters may describe the same event from different perspectives. The limited space of newspapers can also affect how much details about an event can be included in a news article. Newspapers also carry editorials that are usually highly subjective to the political or religious views of the newspaper publishers or the editorial staff. For example, liberals and conservatives have very different opinions on many things. The diversity of opinion becomes even more evident when newspapers from different countries are considered.

Many news readers are interested in knowing different opinions and perspectives about many important events to have a more objective and comprehensive understanding about them. Therefore, the diversity of the news sources that are covered by a news search system can be an important criterion on the quality of the system.

In this paper, the diversity of a search system is defined based on the number of unique countries the retrieved relevant news items come from. Specifically, the following formula is used:

Diversity = # unique source countries/total# of relevant results

In the future, we plan to study other diversity measures.

4.5 Information Richness

The richness here refers to the information contained in the search result records (SRR) returned by a search system. Some information that may be contained in an SRR includes the name of the news source, the size of the article, the URL to the full document, the publication date/time, a short summary/excerpt of the full document, etc. In general, if the SRR from a search system provides all of the above information, it performs well in information richness. We also use a subjective measurement, which is whether the information provided is considered to be sufficient to allow the users to make an informed decision about whether or not to display the full document.

5. EVALUATION AND COMPARISON

In this section, we report our evaluation and comparison of the three news search systems described in Section 3 using the criteria/scheme introduced in Section 4. We will start with the dataset we used for this evaluation.

5.1 Dataset

We used 40 queries in total. They were manually selected by a computer science graduate student who is not involved in the development of AllInOneNews. These queries cover 5 different areas to provide a wide coverage as shown below:

- US news (6)
- International news (13)
- Sports news (5)
- Economy news (10)
- Entertainment news (6)

The queries were taken from the corresponding page of well-known newspapers and news websites such as New York Times, CNN.com and BBC.com from June 13, 2006 to October 30, 2006. The average number of words of these queries is 3.42. In order to make the relevance assessment more objective, whenever a query is generated, a description that explains what the query is about is recorded too. The description is used to judge whether a certain document is relevant or not. All the queries were submitted and judged by the same person so consistency is assured. An example query record is shown in Table 1. It includes the query, the date when the query was submitted to the three news search systems (10.6.06 stands for October 6, 2006 in Table 1), and the description. The full list of queries can be found in the Appendix at the end of this paper.

Table 1. An example query and its description

Query	Date	Description
Peter Pan sequel	10.6.06	A sequel to children's classic Peter Pan has been published - more than 100 years after the original.

For each query the top 10 retrieved documents from each search system are saved. For each retrieved document, its relevance is determined based on its match with the description of the query.

The time when each retrieved document was published online is recorded as accurately as possible. Most news websites publish their news with a date and time. For those that only provide a date, we assume they are published at 12:00am of that day. For a given document we first find out from which country it was published and the time zone of that country so we can transform the local time when the news was published into a standard time (GMT) for this study.

5.2 Experimental Results

We now report the evaluation results based on each criterion described in Section 4.

5.2.1 Traditional effectiveness

Table 2 shows the average TSAP@N values for N = 2, 5, 8 and 10 over the 40 test queries for the three news search systems. We can see that AllInOneNews has the highest TSAP values among the three news search system for N = 5, N = 8 and N = 10 while Google News is the best for N = 2. Mamma News has the lowest values for all Ns. More detailed analysis shows that Google News has the best performance for its top-ranked results (all of its top-ranked results are relevant) while AllInOneNews retrieved more relevant results for N ≥ 2 (please also see the next paragraph).

It may be argued that the TSAP measure puts too much emphasis on higher ranked relevant results. For example, if the top-ranked result is relevant, its score (r_1 = 1.0) is higher than the sum of the scores of the 2nd and the 3rd relevant results ($r_2 + r_3$ = 1/2 + 1/3 = 0.83) and higher than the total of the scores of next six relevant results ($r_4 +\ldots+ r_9 < 1.0$). One way to remedy this is to multiply TSAP@N by R/N [8], where R is the number of relevant results among the top N results. Table 3 compares the performance of the three search systems based on this measure. As it can be seen, AllInOneNews consistently performed better than Google News based on this measure.

Table 2. Average TSAP@N values

N	AllInOneNews	Mamma	Google
2	0.675	0.619	0.688
5	0.406	0.350	0.383
8	0.295	0.247	0.266
10	0.251	0.207	0.222

Table 3. Average TSAP@N * R/N values

N	AllInOneNews	Mamma	Google
2	0.656	0.572	0.625
5	0.379	0.299	0.306
8	0.261	0.195	0.189
10	0.215	0.155	0.147

We would like to point out that the semantic match capability of AllInOneNews was added to the system on October 1, 2006. Among the 40 queries used in this study, only 7 were collected after the above date.

5.2.2 Time-sensitive effectiveness

Table 4 shows the average precision values at different recall levels for the three search systems at cutoff 10 when p = 0.5 using the scheme described in Section 4.2. Figure 1 shows the graph representation of Table 4. We can see that AllInOneNews has an overall much better performance than Mamma News and Google News for this measure. Google News has higher precisions than Mamma News at the two ends of the recall values but Mamma News does better than Google News at recalls 0.4 and 0.5.

Table 4. Average precisions incorporating time-sensitivity

Recall level	Precision		
	AllInOneNews	Mamma	Google
0	0.635	0.403	0.521
0.1	0.628	0.380	0.510
0.2	0.536	0.333	0.381
0.3	0.438	0.246	0.269
0.4	0.337	0.225	0.138
0.5	0.240	0.112	0.073
0.6	0.145	0.062	0.061
0.7	0.122	0	0.041
0.8	0.059	0	0
0.9	0.010	0	0
1.0	0	0	0
Average	0.286	0.16	0.181

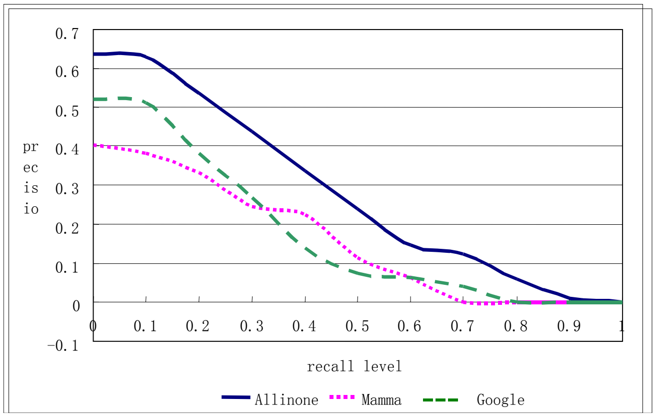

Figure 1. Time-sensitive Effectiveness Comparison

5.2.3 Redundancy

Table 5 shows the average degree of redundancy of the three news search systems we tested. The result shows that Google News has done an impressive job as it managed to eliminate all the redundant results. In contrast, Mamma News has the highest redundancy rate (> 7 redundant results in every 100 results), which is close to twice the redundancy AllInOneNews has.

Table 5. Average degree of redundancy

AllInOneNews	Mamma	Google
3.95%	7.37%	0

5.2.4 Diversity (comprehensiveness)

Table 6 shows the average diversity of each news search system. Google News has the highest diversity among the three systems, followed by AllInOneNews while Mamma News trails the others. It appears that diversity is related to the number of news sources covered. Google News uses 4,500 sources while AllInOneNews connects to over 1,000 news sites. Mamma News probably uses less than 1,000 sources.

Table 6. Average diversity

AllInOneNews	Mamma	Google
0.34	0.24	0.42

5.2.5 Information Richness

Table 7 lists the information that is provided in the search result records by each of the three news search systems.

Although AllInOneNews and Mamma News do not provide the response time that it takes them to search and return results, it is easy to observe that they both take substantially longer time than Google News. This is expected as Google News searches its own index while the other two are metasearch engines that need to send their queries to other search engines and wait for the results to come back.

5.3 Discussions and Comparisons

Although the experimental data shows that AllInOneNews has higher precision at more recall levels, given the small number of queries used, it may seem imprudent to say that AllInOneNews is superior in news search. However, during our investigation, we did observe a number of strengths and weaknesses of these news search/metasearch engines, which are discussed below.

1. Google News maintains a centralized index (possibly replicated) to the news articles from 4,500 news sources. To process a query, Google News needs only to search its own index. In comparison, as indicated earlier, the two news metasearch engines AllInOneNews and Mamma News need to send a user's query to some (or all) of the component newspaper search engines and to receive and merge the results from them before the query results can be presented to the user. Thus, for these two news metasearch engines, more time is required to process a query. For Google News, the average response time for a query is remarkable and seldom takes more than 2 seconds. For Mamma News and AllInOneNews, the average response times are about 4 and 6 seconds, respectively. It is known that Google is powered by over 100,000 computers although it is not clear how much of the resources are devoted to Google News. AllInOneNews is presently powered by a rented server with one dual-processor computer running on MS Windows OS. We did not find information about the server for Mamma News.

Table 7. Information richness

		Description	**AllInOneNews**	**Mamma**	**Google**
Response Time		If the query processing time provided	No	No	Yes
SRR	**Source name**	If the name of the source is included	Yes	Yes	Yes
	# of SRRs	# of SRRs displayed on a page	All retrieved records	20	10
	Description	How useful is the site description *	Very	Medium	Very
	File size	If the size of the retrieved document is indicated	No	No	No
	Date /time	If the SRR indicate a date and time	Yes	Date without time	Recent ones have date and time
Query terms		If they are highlighted in SRRs	Yes	No	Yes

* The usefulness of a site description is a subjective measure. It includes how useful the description lets users make decisions about whether or not to display the full document. Here we find that AllInOneNews and Google News have more detailed description about each SRR so they help users judge if the document is what they want. On the contrary Mamma only provides a shorter description about each SRR.

2. For queries that do not contain exactly the same words used in the news articles for a certain event, Google News often does not return any result. For example, on November 18, 2006, we submitted the query “Moshe Katzav sexual offence” to all three search/metasearch engines. We wanted to find information about the rape allegations against Israeli President Moshe Katzav. AllInOneNews returned several recent relevant news articles. Neither Mamma News nor Google News returned any result for this query. However, Google News did return relevant news if the word “offence” in the query was changed to “offences” or “offense”. Another example is the query “N. Korea rocket launches” submitted on November 18, 2006. Both Mamma News and Google News returned no results while AllInOneNews returned several recent relevant news articles. Google News would return results if the query word “rocket” was changed to “missile”.

3. In general, Google News either does not return any results or returns no good results for old news searches. For example, the query “Ronaldo return Manchester United” which was submitted on July 18th, 2006. It was intended to find news about Cristiano Ronaldo returning to training in Manchester United. Seven relevant results were returned from Google News on July 18th, 2006, but when we did the search again on November 16th, 2006, no relevant results were available. Relevant news articles were returned from AllInOneNews and Mamma News. Previously, Google News mentioned that it keeps news items for 30 days only. Even though this statement has been removed from the Google News site, our test indicates that this practice of keeping news items for only 30 days is continued. We should note that Google has recently launched a separate service for searching archive news.

4. It seems that Mamma News indexes less news from some regions of the world other than US. For example, on July 30th, 2006, Australian Prime Minister John Howard announced that he would contest elections for a fifth term in office. The query “Australia PM fifth term” was submitted on the same day to all three news search/metasearch engines. The top 10 news articles returned from AllInOneNews were all relevant and Google News had 5 relevant results among its top 10 results. However, Mamma News did not return any result. Another example is the news about Google disposing of its 2.6% holding in Baidu. For the query “Google sold Baidu stake” submitted on June 28, 2006, the top 10 results returned by AllInOneNews and Google are relevant, but Mamma News did not return any result.

5. Another interesting observation about Mamma News is that it seems to cover less news sources from the sports and entertainment areas. For example on July 27, 2006, we submitted a query “Landis positive drugs test” which is about Tour de France winner Floyd Landis giving a positive drug test. Among the top 10 results we checked that were

returned from each of the three systems, AllInOneNews gave 10 relevant results and Google gave 9, but Mamma News only returned one. On August 5, 2006 the query "911 movie premieres", which is about Oliver Stone's controversial film "World Trade Center" having a world premiere in New York City, was submitted and Mamma News did not return any documents.

6. The results returned from Mamma News often are not sorted in the degree of relevance to the submitted query. Quite frequently, the top few news articles are not relevant. For example, for the query "cowboys defeat panthers" submitted on October 30th 2006, the second and third news articles returned from Mamma News were not relevant. For some unknown reason, Mamma News occasionally performed very poorly. On October 30th 2006, the query "St. Louis most dangerous" was submitted. Most news articles returned from Google News and AllInOneNews were relevant. However, no results were returned from Mamma News for this query.

7. The experimental results reported in Table 4 suggest that AllInOneNews is more capable of getting fresher news than both Google News and Mamma News. This is consistent with our experience unrelated to this particular comparative study. We believe metasearch engines have the inherent advantage of being able to retrieve more up-to-date information than search engines because the former does not have the delay caused by crawling web pages (news items in this case). The reason why Mamma News did not do well in this regard probably has something to do with its result-merging algorithm. As noted above (item 6), Mamma News is not very good in ranking relevant results higher, which can lead to poor score in time-sensitive effectiveness because this measure considers both freshness and effectiveness. In other words, a news search system can return very fresher results but still gets low score if the fresher news items are not relevant.

8. Based on our test, we believe neither Google News nor Mamma News support semantic match based retrieval. For example, for query "renal calculus" submitted on November 18, 2006, Google News and Mamma News did not return any results while AllInOneNews returned many relevant news items. Supporting semantic match appears to be a unique feature of AllInOneNews.

6. CONCLUSIONS

In this paper, the development of AllInOneNews news metasearch engine is introduced. AllInOneNews has many novel features that are not seen in other metasearch engines such as highly sophisticated search engine selection technique, semantic-based match, and high degree of automation, among others. The successful development and deployment of AllInOneNews shows that it is possible to build large-scale metasearch engines that are practically useful. However, our experience indicates that highly automated solutions are needed to build them cost-effectively and there are still significant technical challenges in improving the robustness of the automated construction tools. Another difficulty we encountered during the development of AllInOneNews is that some news sites do not have search engines or have poor quality search engines. One way to solve this problem is to build search engines for these sites and use them as component search engines.

We also evaluated and compared three news search systems (Google News, Mamma News, and AllInOneNews) based on five criteria (traditional effectiveness, time-sensitive effectiveness, degree of redundancy, diversity, and information richness) using 40 queries. The results showed that AllInOneNews has some advantage on effectiveness (including time-sensitive effectiveness) over Google News and Mamma News, while Google News is clearly better in removing redundant results and providing better diversity than the two metasearch engines.

In this paper, we introduced a new precision measure for the evaluation of systems that retrieve time-sensitive information and devised a scheme to compare multiple search systems using this measure. In addition to the relevance of a piece of retrieved information, this new measure takes into account the recency of the information.

We should point out that the experimental data were collected based on only 40 queries. As such, it may not be conclusive to say that AllInOneNews definitely has better retrieval performance than the other two news search systems. We plan to do more extensive evaluation in the future. Nevertheless, we have identified several strengths and weaknesses of each search/metasearch engine. Although AllInOneNews showed better performance, it requires longer processing time. Google News has the best response time but performs poorly for certain types of queries. For Mamma News, the relevant news articles are often not the top most ones.

7. ACKNOWLEDGMENTS

This work is support in part by the following SBIR grants from NSF: DMI-0340348 (Phase I) and DMI-0522271 (Phase II).

8. REFERENCES

[1] C. Baumgarten. A probabilistic solutions to the selection and fusion problem in distributed information retrieval. ACM SIGIR Conference, 1999.

[2] M. Bergman. The Deep Web: Surfacing Hidden Value. White Paper of CompletePlanet at http://brightplanet.com/pdf/deepwebwhitepaper.pdf, 2001.

[3] L Barbosa, J Freire. Searching for hidden-web databases. 8th International Workshop on WebDB, 2005.

[4] J. Callan, Z. Lu, and. W. Croft. Searching Distributed Collections with Inference Networks. ACM SIGIR, 1995, pp.21-28.

[5] J. Cope, N. Craswell, D. Hawking. Automated Discovery of Search Interfaces on the Web. ADC 2003: 181-189

[6] D. Dreilinger, and A. Howe. Experiences with selecting search engines using metasearch. ACM Transactions on Information Systems, July, 1997, pp.195-222.

[7] Y. Fan, and S. Gauch. Adaptive Agents for Information Gathering from Multiple, Distributed Information Sources. 1999 AAAI Symposium on Intelligent Agents in Cyberspace, Stanford University, March 1999.

[8] S. Gauch, G. Wang, and M. Gomez. ProFusion: Intelligent fusion from multiple, distributed search engines. Journal of Universal Computer Science, 1996.

[9] L. Gravano, and H. Garcia-Molina. Generalizing gloss to vector-space databases and broker hierarchies. VLDB, 1995, pp.78-89.

[10] L. Gravano, and H. Garcia-Molina. Merging ranks from heterogeneous Internet sources. VLDB, 1997, pp.196-205.

[11] D. Hawking, N. Craswell, and K. Griffiths. Which search engine is best at finding online services? WWW conference, poster, 2001.

[12] D. Hawking, N. Craswell, P. Bailey, K. Griffiths. Measuring Search Engine Quality. Information Retrieval, 4(1), 2001.

[13] K.L. Liu, C. Yu, W. Meng, W. Wu, and N. Rishe. A Statistical Method for Estimating the Usefulness of Text Databases. IEEE TKDE, 2002.

[14] Y. Lu, W. Meng, L. Shu, C. Yu, and K.L. Liu. Evaluation of Result Merging Strategies for Metasearch Engines. WISE Conference, pp.53-66, November 2005.

[15] Y. Lu, W. Meng, W. Zhang, K.L. Liu, and C. Yu. Automatic Extraction of Publication Time from News Search Results. Int'l Workshop on Challenges in Web Information Retrieval and Integration (WIRI2006), April 2006.

[16] U. Manber, and P. Bigot. The Search Broker. USENIX Symposium and Internet Techniques and Systems, Monterey, California, December, 1997, pp.231-239.

[17] W. Meng, K.L. Liu, C. Yu, X. Wang, Y. Chang and N. Rishe. Determining Text Databases to Search in the Internet. VLDB, 1998.

[18] W. Meng, Z. Wu, C. Yu, and Z. Li. A Highly-Scalable and Effective Method for Metasearch. ACM Transactions on Information Systems 19(3), pp.310-335, July 2001.

[19] W. Meng, C. Yu, and K.L. Liu. Building Efficient and Effective Metasearch Engines. ACM Computing Surveys, 34(1), March 2002, pp.48-84.

[20] Y. Rasolofo, D. Hawking, and J. Savoy. Result merging strategies for a current news metasearcher. Information Processing & Management, 39, 2003, pp.581-609.

[21] Z. Wu, W. Meng, C. Yu, and Z. Li. Towards a highly scalable and effective metasearch engine. WWW Conference, Hong Kong, 2001.

[22] Z. Wu, V. Raghavan, H. Qian, V. Rama K, W. Meng, H. He, and C. Yu. Towards Automatic Incorporation of Search Engines into a Large-Scale Metasearch Engine. IEEE/WIC International Conference on Web Intelligence, 2003.

[23] C. Yu, W. Meng, K.L. Liu, W. Wu and N. Rishe. Efficient and Effective Metasearch for a Large Number of Text Databases. ACM CIKM, November 1999.

[24] C. Yu, K. Liu, W. Meng, Z. Wu, and N. Rishe. A Methodology to Retrieve Text Documents from Multiple Databases. IEEE TKDE, Vol.14, No.6, November/December 2002, pp.1347-1361.

[25] C. Yu, and W. Meng. Web Search Technology. In *The Internet Encyclopedia* edited by Hossein Bidgoli, Wiley Publishers, pp.738-753, 2003.

[26] B. Yuwono, and D. Lee. Server Ranking for Distributed Text Resource Systems on the Internet. DASFAA, 1997, pp.391-400.

[27] H. Zhao, W. Meng, Z. Wu, V. Raghavan, and C. Yu. Fully Automatic Wrapper Generation for Search Engines. WWW Conference, pp.66-75, 2005.

Appendix: Queries Used in Evaluation

World News

Query / Time	**Description**
Thailand coup / 9.20.06	The leader of the military coup in Thailand has said a new prime minister will be named within two weeks.
China Tibet railway / 6.30.06	The highest - and most controversial - railway in the world begins operating on Saturday between China and Tibet.
Bush visit Baghdad / 6.13.06	President Bush arrived today in Baghdad on an unannounced visit to meet with Iraqi Prime Minister Nuri al-Maliki. It is his first visit since Iraq's permanent government was approved May 20. Bush and his top officials began a summit at Camp David on Monday to discuss Iraq's progress and scheduled a teleconference with al-Maliki for today.
Typhoon Kaemi China / 7.25.06	Typhoon Kaemi has reached China's south-eastern coast, bringing with it heavy rain and high winds.
Europe heatwave death / 7.19.06	A heatwave affecting much of Western Europe has resulted in several deaths, health officials say.
Japanese PM continue shrine visits / 7.20.06	Japanese premier Junichiro Koizumi says he will continue visiting a controversial Tokyo war shrine despite evidence the former emperor opposed it.
Hezbollah rockets Haifa / 7.16.06	JERUSALEM - Saying it was responding to overnight Israeli airstrikes inside Lebanon, the militant group Hezbollah claimed responsibility Sunday for pounding the northern Israeli city of Haifa with rockets, killing eight Israelis and prompting new airstrikes in southern Beirut.
taiwan president second recall / 10.13.06	A second attempt to pass a recall motion against Taiwanese President Chen Shui-bian has failed.
Ethiopia flood / 8.17.06	Search and rescue teams are scouring flood waters in southern Ethiopia as bad weather continues to hamper a round-the-clock hunt for survivors.
Toronto HIV conference / 8.13.06	Twenty-four thousand delegates are gathering for a big international conference on HIV and Aids, which begins in Toronto on Sunday.
North Korea missile test / 7.5.06	World powers have condemned North Korea for test-firing a series of missiles, including one thought capable of reaching the US.
Indonesia bird flu / 8.8.06	Two Indonesian teenagers have died of the bird flu virus, bringing the country's number of human fatalities to 44, Indonesian health officials say.
Australia PM fifth term 7.31.06	Australian Prime Minister John Howard has announced that he will contest elections for a fifth term in office.

U.S. news

Query / Time	**Description**

Chicago apartment fire / 9.3.06	CHICAGO - Six children, including two 3-year-olds, died in a fire started by candles that family members used because their apartment lacked electricity, fire commissioner Raymond Orozco said on Sunday.
US population 300 million / 10.17.06	US population reaches 300 million
Mexico fence law / 10.26.06	US President George W Bush has signed into law a plan for 700 miles (1,125km) of new fencing along the US-Mexico border, to curb illegal immigration.
US soldier murder charges Iraq / 7.4.06	A former US soldier has appeared in a US federal court, charged with the rape and murder of an Iraqi woman, and the killing of three members of her family.
Petrol prices US production / 6.15.06	US industrial production unexpectedly fell in May, highlighting weaknesses in manufacturing and signaling that the world's largest economy may be cooling. Industrial output dropped 0.1% in May after April's 0.8% rise, the Federal Reserve reported.
St. Louis most dangerous / 10.30.06	A surge in violence made St. Louis, Missouri, the most dangerous city in the nation, according to an annual list by Morgan Quitano Press based on FBI figures.

Sports News

Query / Time	Description
Dusty Baker fired / 7.7.06	In the always-intriguing annual race to be the first manager to the unemployment line, Dusty Baker is leaving everyone in the -- pardon the expression -- dust. Baker's firing appears inevitable
Landis positive drugs test / 7.27.06	Tour de France winner Floyd Landis has given a positive drugs test, according to his Phonak team.
yankees beat red sox / 8.19.06	The Yankees beat the Boston Red Sox 14-11 to complete a sweep of Friday's day-night doubleheader and give New York a season-high 3 1/2-game lead in the division.
Ronaldo return Manchester United / 7.18.06	Manchester United boss Sir Alex Ferguson expects Cristiano Ronaldo to return to training on 31 July and insists the winger is going nowhere.
cowboys defeat panthers / 10.30.06	NFL: The Carolina Panthers was defeated 35:14 by Dallas Cowboys.

Business News

Query / Time	Description
Airbus orders fall behind Boeing / 7.6.06	European aircraft maker Airbus has fallen behind arch US rival Boeing in the number of new orders for planes.
Ryanair complain Air France / 6.13.06	Budget airline Ryanair has filed a further complaint with the European Commission accusing rival Air France KLM of trying to block competition.
Google sold Baidu stake / 6.28.06	The US company confirmed that it had disposed of its 2.6% holding in Baidu - acquired before the latter's 2005 stock market flotation - on Wednesday.
NetIDMe card / 8.2.06	A virtual ID card designed to improve children's net safety has been launched in the UK, US, Canada and Australia.
china cut export rebates / 7.23.06	China is considering cutting tax rebates on some of its exports in order to tackle its record trade surplus, state media have reported.
Yukos bankrupt / 8.1.06	Troubled Russian oil firm Yukos has been declared bankrupt by a court in Moscow, clearing the way for the firm to be liquidated.
honda plane 2010 / 8.9.06	Honda is setting up a new US business to oversee the production of its mini passenger jet and says it plans to launch the plane in 2010.
dell recall laptop batteries / 8.15.06	The world's largest manufacturer of personal computers, Dell, is to recall 4.1 million of its notebook computer batteries because of a fire risk.
Saudi Arabia buys Eurofighters / 8.18.06	Saudi Arabia has confirmed it is to buy 72 Eurofighter Typhoon aircraft from the UK, in a deal that could be worth more than 6bn.
windows 98 shut down / 7.11.06	Microsoft is urging an estimated 70 million users of Windows 98 to upgrade as it ends support for the software.

Entertainment News

Query / Time	Description
911 movie premieres / 8.5.06	Oliver Stone's controversial film World Trade Center has had its world premiere in New York.
Arthur Lee dies / 8.4.06	Arthur Lee, singer and guitarist of the influential 1960s band Love, has died in Memphis at the age of 61 following a battle with acute myeloid leukaemia.
Mel Gibson drink-driving / 7.30.06	Hollywood actor and director Mel Gibson has said he is "ashamed" of the actions that led to his arrest for drink-driving early on Friday morning.
Peter Pan sequel / 10.6.06	A sequel to children's classic Peter Pan has been published - more than 100 years after the original.
Martin Scorsese quit Hollywood / 10.16.06	Film director Martin Scorsese says he plans to take a break from Hollywood to make low-budget films.
South Park Steve Irwin / 10.27.06	The cartoon series South Park's latest episode has caused outrage by featuring the recently deceased Crocodile Hunter, Steve Irwin.