Received 8 May 2006
Revised 28 August 2006
Accepted 30 August 2006

Improving the search process through ontology-based adaptive semantic search

Chyan Yang, Keng-Chieh Yang and Hsu-Chieh Yuan
Institute of Information Management, National Chiao Tung University, Taipei, Taiwan

Abstract

Purpose – The purpose of this research is to describe an efficient search methodology to help improve the search results in the top portion of a lengthy search list. When facing a lengthy search list, people often limit themselves to the top ten items on the list, even though there may be more useful information after the top ten items.

Design/methodology/approach – This study proposes an ontology-based adaptive semantic search to significantly improve the search experience. To capture the semantic difference of search terms, naïve ontology is used to store the relationship among terms. Before a search term is processed by the search engine Lucene, the related words of the search term are selected from ontology structures to form new query phrases in the process of query expansion. The weighting of the expanded query phrases is dynamically learned by observing the users' clicking behaviors.

Findings – Research results show that with the aid of ontology the average precision rate of all cases is dramatically higher than the precision rate for the default search result. Even in the worst cases, in some situations, this ontology is still close to the precision rate for the default search result.

Originality/value – This paper shows how it is possible to improve the precision rate of items retrieved after a search and thus avoid information overload.

Keywords Information retrieval, Semantics, Information searches, Search engines

Paper type Research paper

Introduction

The more resources there are available in the form of electronic documents, records, and reports on the Internet, the more people feel overwhelmed by the amount of information they have at hand. Information overloading has become a chronic headache for every knowledge worker, resulting in a demand for advanced information retrieval techniques to process data in such a way that can separate the valuable information from millions of less relevant documents. Thus, document searching, clustering, and summarization have been active research fields for decades that mitigate the threat of information overloading.

The role of search is at least as essential as the role that content was in the last century. In the past, people believed that content was king, which implied the challenge of how to get the information they needed. As the scarceness of time is more often a concern than the scarceness of information, it now turns out that search, not content, is king. Most of the time people know that useful information is out there, but find it difficult to access the right information in a quick, easy, and painless fashion.

The quality of search results has indeed become more important than the quantity of search results. More specifically, the top ten listings of search results are what the

The Electronic Library
Vol. 25 No. 2, 2007
pp. 234-248
© Emerald Group Publishing Limited
0264-0473
DOI 10.1108/02640470710741359

vast majority of people actually look at during a search according to an eye tracking study conducted by search marketing firms Enquiro and Did-it and eye tracking firm Eyetools (Enquiro Search Solutions Inc., 2005). Figure 1 shows that most eye tracking activity during a search take places in a triangle at the left of the top of the search results, representing areas with maximum interest – called a golden triangle – for

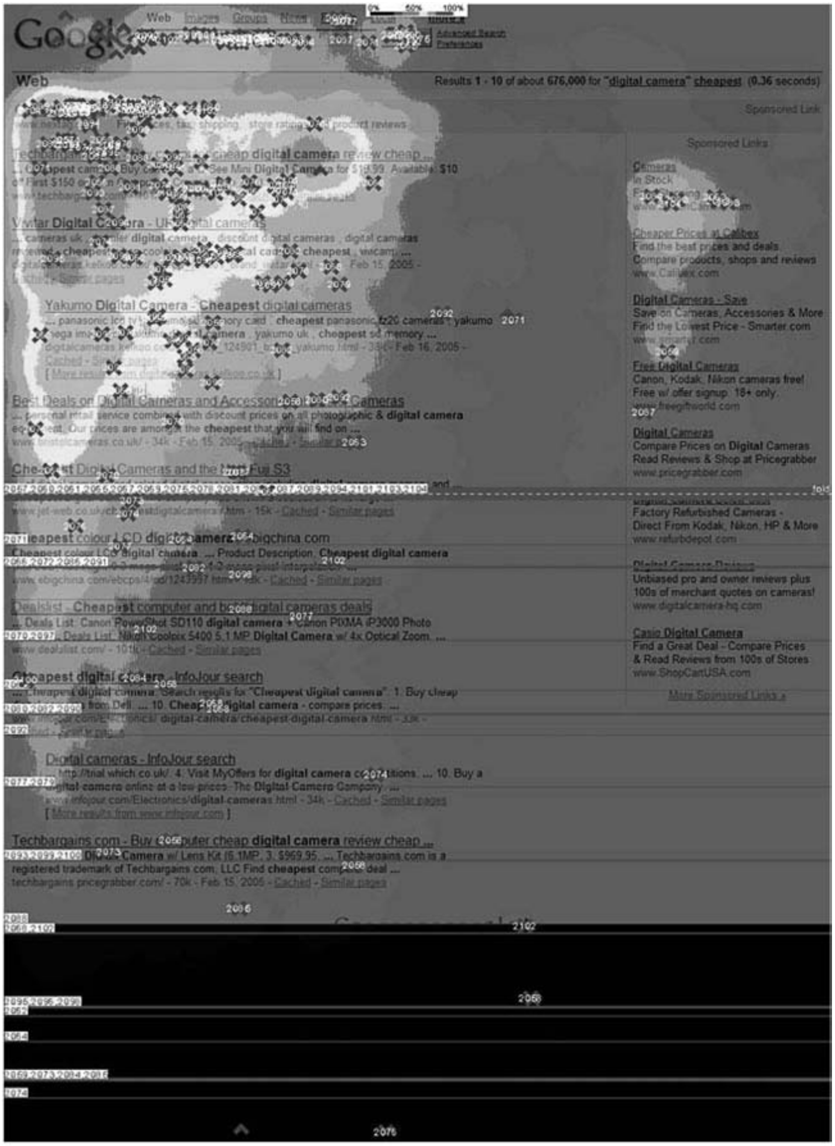

Web
Images
Groups
News
Advanced Search
Preferences
Results 1 - 10 of about 676,000 for "digital camera" cheapest (0.36 seconds)
Sponsored Link
Sponsored Links
Vivitar Digital Camera - UK digital cameras
Yakumo Digital Camera - Cheapest digital cameras
Best Deals on Digital Cameras and Accessories
Dealslist - Cheapest computer and best digital cameras deals
Cheapest digital camera - InfoJour search
Digital cameras - InfoJour search
Techbargains.com - Buy computer cheap digital camera review cheap ...
Cheaper Prices at Calibex
Digital Cameras - Save
Free Digital Cameras
Digital Cameras
Casio Digital Camera
More Sponsored Links »
fold

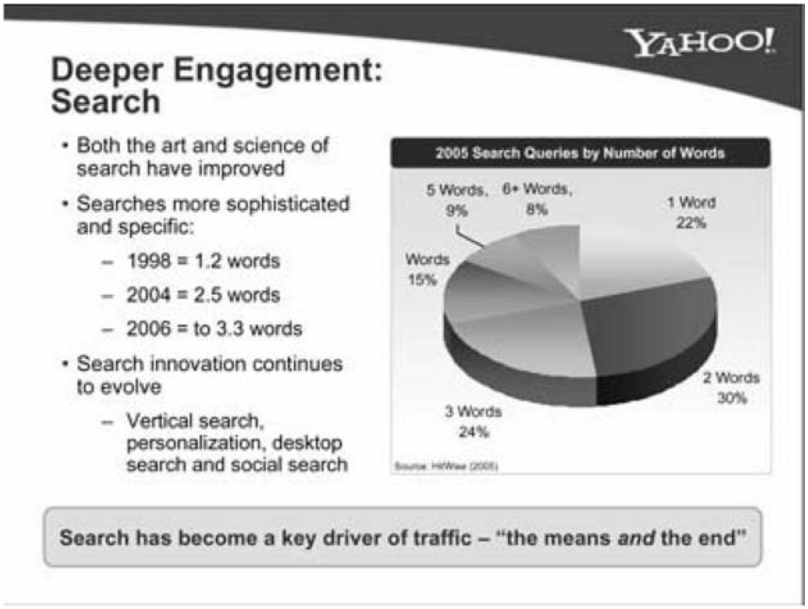
Figure 1.
Search's Golden Triangle

Source: Enquiro (2005)

marketing experts. In the study, 50 percent of participants looked at only the top seven listings of the search results, and only 20 percent of participants looked at all top ten of the search results.

Another study of search engine usage in North America in 2004 (Enquiro Search Solutions Inc., 2005) suggested that less than 25 percent of respondents frequently use advance search features on a search engine, and that over 70 percent of respondents input general words in the search box. Pu and Yang (2003) explored the possibility of adding user-oriented class associations to hierarchical library classification schemes. The results showed that classification schemes can be made more adaptable to changes of users and the uses of different library collections by analyzing the circulation patterns of similar users. Yang *et al.* (2006) developed an IT specification extraction system and examined the specifications of IT products with various brands and patterns. This system combined the natural language process with ontology concept. The results showed that the system can be effective for IT purchasing applications. The fact that most search queries are 1.2 words in 1998, 2.5 words in 2004, and 3.3 words in 2006 as shown in Figure 2 (Bogatin, 2006). Bogatin's findings are also confirmed by a list of popular queries for 2005 on Google (www.google.com/intl/en/press/ zeitgeist2005.html) and on Yahoo! (http://tools.search.yahoo.com/top2005/) as shown in Figure 3 and Figure 4, respectively.

The way people search implies that the precision rate is more important than the recall rate. Therefore, the recall rate is not measured in the experiments. Furthermore, only the precision rate of the top twenty search results is used to evaluate the performance of our system. Considering the fact that if we measure the precision rate of the top ten listings of the search results, the change of one matched document would result in a 10 percent point change in the precision rate. As such, we choose to measure



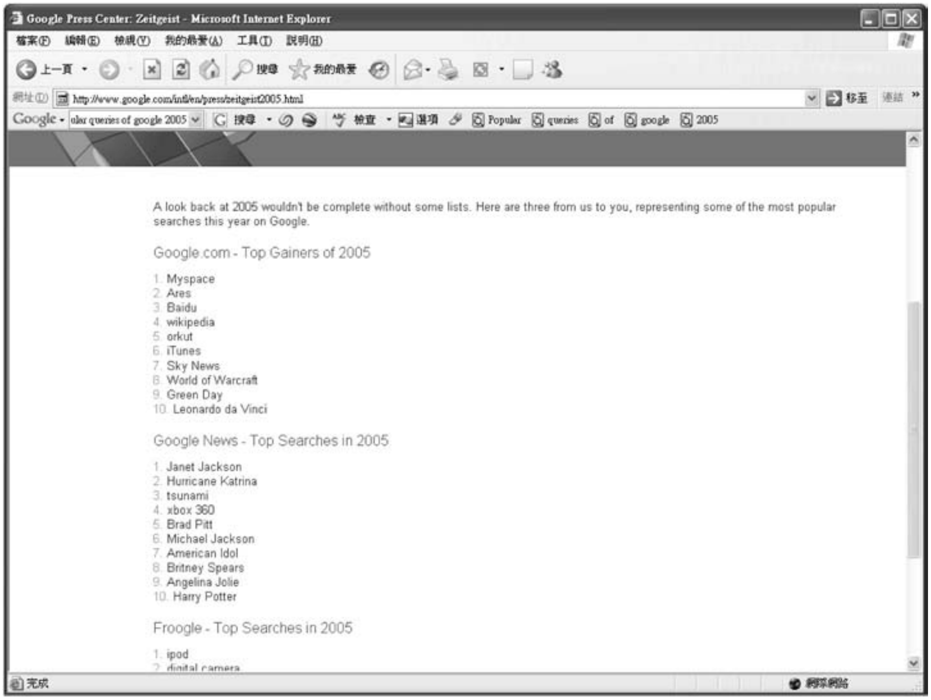
YAHOO!

Deeper Engagement: Search

- Both the art and science of search have improved
- Searches more sophisticated and specific:
  - 1998 = 1.2 words
  - 2004 = 2.5 words
  - 2006 = to 3.3 words
- Search innovation continues to evolve
  - Vertical search, personalization, desktop search and social search

Search has become a key driver of traffic – "the means *and* the end"

Figure 2.
Yahoo 2005 search queries by number of words


Google Press Center: Zeitgeist - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁
搜尋
我的最愛
網址(D)
http://www.google.com/intl/en/press/zeitgeist2005.html
移至
連結
Google
ular queries of google 2005
搜尋
檢查
選項
Popular
queries
of
google
2005
完成
網際網路

A look back at 2005 wouldn't be complete without some lists. Here are three from us to you, representing some of the most popular searches this year on Google.

Google.com - Top Gainers of 2005

1. Myspace
2. Ares
3. Baidu
4. wikipedia
5. orkut
6. iTunes
7. Sky News
8. World of Warcraft
9. Green Day
10. Leonardo da Vinci

Google News - Top Searches in 2005

1. Janet Jackson
2. Hurricane Katrina
3. tsunami
4. xbox 360
5. Brad Pitt
6. Michael Jackson
7. American Idol
8. Britney Spears
9. Angelina Jolie
10. Harry Potter

Froogle - Top Searches in 2005

1. ipod
2. digital camera

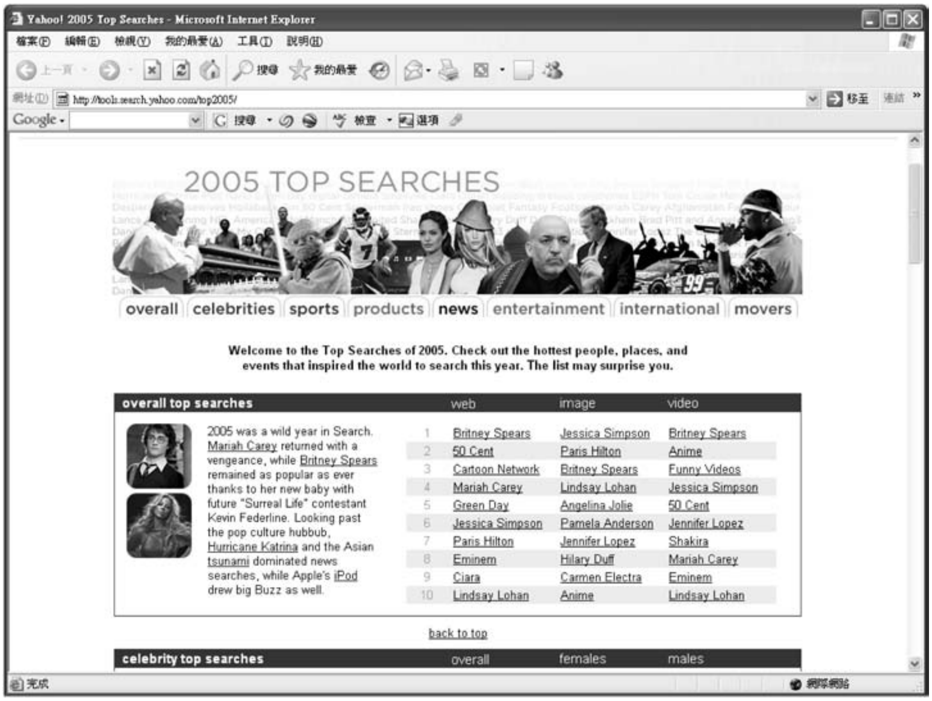
Figure 3.
Popular queries for 2005 on Google

the top 20 search results instead of the top ten in the hope for smaller changes in the precision rate and more detailed performance results.

The fact that most searches are conducted based on 2.5 words (Bogatin, 2006) calls for a structural way of organizing the relationship among words. We propose a naïve ontology structure which is used to capture the semantic differences of words and expand the search terms before the search is processed. To dynamically learn a user's preference for ontology structures, a relevance feedback loop is employed to observe the user's clicking behavior and adjust the weight of each ontology structure accordingly.

Information retrieval

Brin and Page (1998), the founders of the gigantic search engine Google at Stanford University, devised a creative way of ranking the importance of each web page by taking into account both the importance of the web pages linked to the web page and the number of links the web page receives. This underlying technique called PageRank is a uniquely democratic system of leveraging the hyperlink structure of the web pages. This hyperlink structure, however, does not always exist in word files, acrobat documents, and presentation slides, making Google less effective for files without hyperlinks than for web pages. Another limitation of using Google is that people around the world who input the same keywords at the same time will get exactly the same search result regardless of their past search history on Google.



Welcome to the Top Searches of 2005. Check out the hottest people, places, and events that inspired the world to search this year. The list may surprise you.

overall top searches

2005 was a wild year in Search. Mariah Carey returned with a vengeance, while Britney Spears remained as popular as ever thanks to her new baby with future "Surreal Life" contestant Kevin Federline. Looking past the pop culture hubbub, Hurricane Katrina and the Asian tsunami dominated news searches, while Apple's iPod drew big Buzz as well.

	web	image	video
1	Britney Spears	Jessica Simpson	Britney Spears
2	50 Cent	Paris Hilton	Anime
3	Cartoon Network	Britney Spears	Funny Videos
4	Mariah Carey	Lindsay Lohan	Jessica Simpson
5	Green Day	Angelina Jolie	50 Cent
6	Jessica Simpson	Pamela Anderson	Jennifer Lopez
7	Paris Hilton	Jennifer Lopez	Shakira
8	Eminem	Hilary Duff	Mariah Carey
9	Ciara	Carmen Electra	Eminem
10	Lindsay Lohan	Anime	Lindsay Lohan

Figure 4.
Popular queries for 2005 on Yahoo!

When Google searches the contents of web pages, it also reckons how frequently a site is linked to and what words are used to describe the hyperlinks to a site. Therefore, a small number of people can influence the results of Google searches by adding links to a site with specific words. For example, web users searching "miserable failure" on Google are led to the biography of USA President George W. Bush on the White House website. This technique is called Google bombing, showing that the PageRank algorithm is still subject to artificial manipulation.

Query expansion

Query expansion has been studied extensively since the early years of information retrieval, because search accuracy largely depends on the quality of search keywords and users seldom use the right search keywords on their first search. In fact, the results of an original query are often unsatisfactory to the users. By adding more related keywords, query expansion transforms the original search based on the existing knowledge structures such as a dictionary and thesaurus. However, this approach does not prove to be universally useful.

Jing and Croft (1994) found that an automatically-generated association thesaurus improves retrieval performance. Voorhees (1994) and Jones *et al.* (1995) analyzed an intellectually-created thesaurus for query expansion and discovered no significant improvement in retrieval performance. The effectiveness of query expansion with domain specific thesaurus was studied by Kristensen (1993) and Jarvelin *et al.* (1996). They found that a thesaurus-based query expansion improves retrieval performance in best match and Boolean cases. However, these studies ignore the interaction among search keywords, query structure, and query expansion. Thus, the usefulness of different kinds of thesauruses cannot be determined without considering the structural characteristics of queries.

Lu and Keefer (1995) analyzed how query size affects retrieval effectiveness by altering TREC queries. They found that reducing the query terms has a negative impact on retrieval performance, and that query expansion with an automatically-generated association thesaurus improves retrieval performance. They provided some evidence that the effectiveness of different retrieval models depend on the number of search keywords in the query. It is believed that the shorter the queries are, the more gain there is from the query expansion (Jing and Croft,1994).

Relevance feedback

Relevance feedback processes can be classified based on two criteria. The first criterion is based on how to obtain relevance information, which can be assigned by the user (Rocchio, 1971) or solely rely on the order of the listings in the preliminary output (Harman, 1988). When it comes to evaluating the systems, this distinction is very important. The second criterion is how to change the original query, which is often closely related to the retrieval model being used.

To avoid imposing a computation burden, systems generally put a limit on the number of documents used to extract feedback terms and the number of extracted terms. Salton and Buckley (1990) pointed out that methods based on the Rocchio algorithm are more effective than probabilistic modes as well as they are more efficient. The generally accepted conclusions of relevance feedback processes are as follows:

- Collections that produce relatively bad output in an initial retrieval can be improved more dramatically than collections that perform well in the initial search.
- The average length of the initial queries largely determines the performance of relevance feedback. Short queries are more applicable to optimization than longer ones.
- An analysis of all the terms in the documents selected for the relevance feedback process performs better than a partial expansion by the highest weighted terms.

Our proposed approach is DirectHit (www.directhit.com/), which is a site ranking system. DirectHit, which had been used by search engines such as Ask, Lycos, and Hotbot, is based on the concepts of click popularity and stickiness of a site (Search Engine Inc., 2003).

The click popularity of a site is measured by the number of clicks it received in search results pages. The stickiness of a site is measured by the amount of the time a user spends at it. It is calculated as the time between each of the user's clicks on the search results pages. Given DirectHit's implicit feedback loop and people's inherent reluctance to actively express their opinions, we prefer DirectHit to other explicit feedback loop mechanisms.

System overview

Corpus is a collection of news reports at www.ettoday.com/ for 2002 and 2003. News reports were chosen to be corpus, because news reports cover different aspects of daily business activities as well as local events. The variety of topics and information makes news corpus quite challenging as document collections in today's business world. There are 109,491 news reports for 2002 and 125,830 news reports for 2003, for a total of 235,321 news reports. Table I and Figure 5 reveal that half of the reports – 50.56 percent – have a size of less than 1,000 bytes (500 Chinese words), whereas there 10.2 percent of the reports have a size larger than 1800 bytes (900 Chinese words). The reports whose size is between 600 bytes (300 Chinese words) and 1600 bytes (800 Chinese words) make up 70.10 percent. All news reports have html tags and irrelevant parts like navigation menus and advertisements removed before being indexed by the search engine Lucene (http://lucene.apache.org/).

Ontology

A naïve format is chosen to represent the relationship among elements in an ontology structure. There are three items in a row separated by a comma: X element, Y element, and the relationship between X and Y. There are three types of relationships: the first type is a super-class relationship, indicating that Y is a super-class of X; the second type is an equivalent relationship, indicating that Y is equivalent to X; the last type is a sub-class relationship, indicating that Y is a sub-class of X.

After going through the business news reports in corpus, we manually edit the Business ontology structures shown in Table II. There are three categories in business ontology structures. Category 1 focuses on news reports regarding stock price; category 2 focuses on news reports that are related with orders and products; category 3 focuses on news reports that are pertinent to doing business in China.

Length (bytes)	Count	%	Cumulative %
0-199	1,520	0.65	0.65
200-399	9,377	3.98	4.63
400-599	23,126	9.83	14.46
600-799	42,059	17.87	32.33
800-999	42,888	18.23	50.56
1,000-1,199	34,790	14.78	65.34
1,200-1,399	26,479	11.25	76.59
1,400-1,599	18,761	7.97	84.57
1,600-1,799	12,309	5.23	89.80
1,800-1,999	7,763	3.30	93.09
2,000-2,199	4,996	2.12	95.22
2,200-2,399	3,259	1.38	96.60
2,400-2,599	2,193	0.93	97.53
2,600-2,799	1,516	0.64	98.18
2,800-2,999	1,024	0.44	98.61
3,000-3,199	723	0.31	98.92
3,200-3,399	533	0.23	99.15
3,400-3,599	377	0.16	99.31
3,600-3,799	289	0.12	99.43
3,800-3,999	209	0.09	99.52
4,000-	1,130	0.48	100.00

Table I.
The length of reports in 2002 and 2003

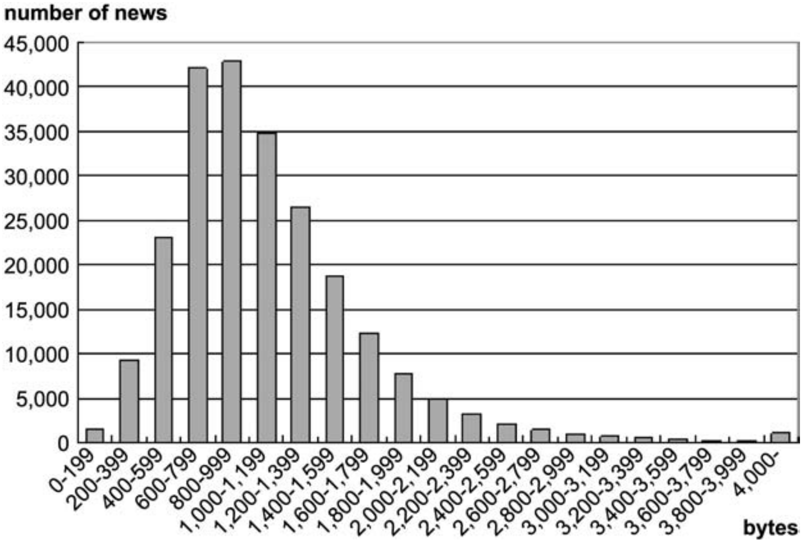


Figure 5.
The length distribution of news reports in 2002 and 2003

Table II.
Keywords for each category in business ontology

Category	Keywords
Category 1	Stock price, stock dividend,(foreign investors, buyers over sellers, sellers over buyers, revenue, profit, profit, gross profit
Category 2	Orders, stealing orders, products, release, product delivery
Category 3	Taiwanese companies, China, communists, mainland China, investment in China, cross straits, Shanghai, Beijing, Hangzhou, Suzhou, Guangzhou, Dongguan, Fujian, Shenzhen, Xiamen

Search engine Lucene

Lucene is an open-source, high-performance text search engine with rich features written in Java originally by Doug Cutting. Initially, it started as an open-source project at the SourceForge website in March 2000. In September 2001 it joined Apache Software Foundation's Jakarta family and in May 2006 version 2.0.0 was released. Many large notable organizations including FedEx, the Mayo Clinic, Hewlett-Packard, Epiphany, and MIT's OpenCourseware use Lucene. However, it requires the user making changes to the source code in order to handle large collections of search requests. Lucene's similarity scoring formula is shown in the following equation and the definitions of factors are described in Table III. The score is computed for each matched document:

$$\sum_{t \text{ in } q} \mathrm{tf}(t \text{ in } d) \cdot \mathrm{idf}(t) \cdot \mathrm{boost}(\mathrm{t.field} \text{ in } d) \cdot \mathrm{lengthNorm}(\mathrm{t.field} \text{ in } d) \cdot \mathrm{Coord}(q,\ d) \cdot \mathrm{queryNorm}(q) \tag{1}$$

Table III.
Factors in Lucene's scoring formula

Factor	Description
tf(*t* in *d*)	Term frequency for the term (*t*) in the document (*d*)
idf(*t*)	Inverse document frequency of the term (*t*)
boost(t.field in *d*)	Field boost value for the field (*t*) in the document (*d*)
lengthNorm(t.field in *d*)	Normalization value for the field (*t*) based on the number of terms in the field (*t*).
Coord(*q*, *d*)	Coordination factor, based on the number of query terms (*q*) the document (*d*) contains
queryNorm(*q*)	Normalization value for a query, given the sum of squared weights of each of the query terms

Lucene supports all kinds of powerful search expressions including Boolean operators, field selection, wildcard, range, proximity, and boost. Among all those expressions, Boolean OR operator, proximity, and boost are used in the following experiments. Proximity expression restricts Lucene to find words that are within a specified distance. To do a proximity search, a " ~ " symbol is appended at the end of a search phrase along with a distance value. Boost expression assigns the relevance level of terms with a boost factor. To boost a term, a "^" symbol with a boost factor is appended at the end of a term. The higher the boost factor is, the more relevant the term will be.

System architecture

The system architecture as shown in Figure 6 can be divided into two parts, which are a query expansion process and a feedback loop. The Naïve ontology structure described above is used to store the semantic differences of words and to expand the input search terms. All of the super-classes, sub-classes, and equivalent classes of the search term are selected from ontology structures. The sub-classes of the sub-classes of the search term also are selected if those classes exist in the structures. Expanded search terms are added with the weight of ontology structures from which they are expanded.

For each query term, the system keeps track of a hit counter for each ontology structure. The weight of an ontology structure is calculated as 100 multiplied by the ratio between the hit count of the search term and the highest hit counter of the search terms. The expanded query is next processed by the search engine Lucene. After Lucene returns the search result, a feedback loop monitors the user's click-through behavior. Once a click is made by the user, all expanded terms are looked up for the selected news item. The ontology with the largest number of occurrences of expanded terms in the news content increases its hit counter for the search term by one.

Experiments with business ontology

The experiment consists of the following steps.

Step 1. Choose one category as the category of interest from unselected categories from the business ontology structures, and repeat step 1 to step 4 three times.

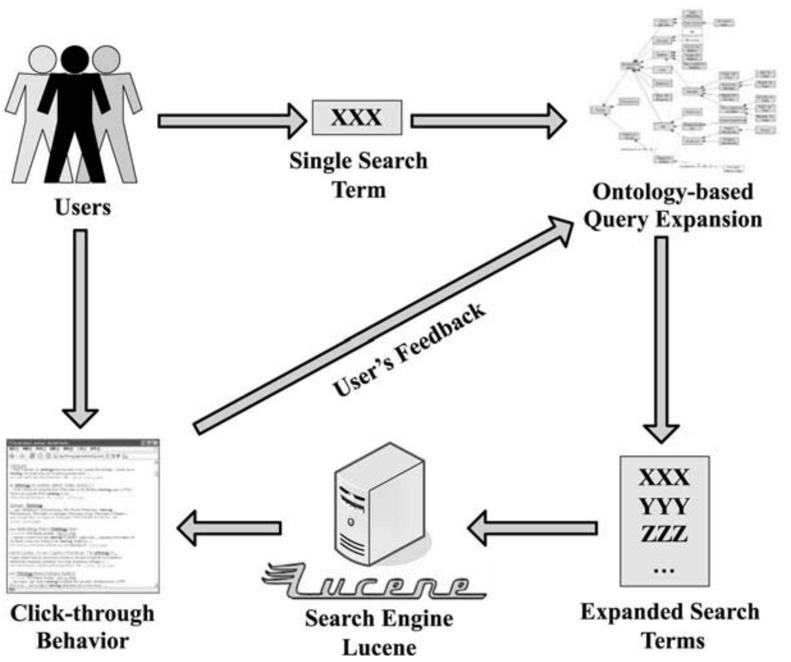


Figure 6.
System architecture

Step 2. Vary the weights of the ontology structures of each categories from 0 to 100 by a unit of 10.

Step 3. Combine each of the elements in categories with the given company name to form a Lucene proximity search phrase with a distance value of 25 along with a boost value that is the weight of the category of the element.

Step 4. Feed the expanded query terms in step 3 into the search engine Lucene and calculate the precision rate of the top twenty listings of the search result with respect to the category of interest.

In order to test the experiment, we choose some well-known corporations in Taiwan which have high occurrences in news corpus, such as China Steel Corporation (CTC), Formosa Plastics Corporation (FPC), Taiwan Semiconductor Manufacturing Corporation (TSMC), Acer Corporation (acer), Asustek Computer Inc. (ASUS), and United Microelectronic Corporation (UMC).

The following tables show the number of news reports that cover the search term in the category on the row and the highest achievable precision rate of the search term in the category on the row when the category on the row is the category of interest during a search.

If the count of the search term in the category on the row is larger than 20, then the highest achievable precision rate of the category on the row is 100 percent. If not, then it is calculated as the count of the search term in the category on the row divided by 20, because we measure the precision rate of only the top twenty search results (see Table IV).

Company	Category type	Count
CTC	1	304
	2	63
	3	85
FPC	1	240
	2	104
	3	143
TSMC	1	1,312
	2	362
	3	304
UMC	1	1,232
	2	298
	3	147
ACER	1	93
	2	185
	3	39
ASUS	1	268
	2	168
	3	38

Note: Highest achievable precision rate are all of 100 percent of each

Table IV.
Statistics for the search of company

The following correlation tables show the correlation among each category. In the correlation tables each item is the percentage of overlap as computed by the document set of category in the row versus the document set of category on the column (see Table V).

The summary tables show the experiment results for each category. In the summary tables each row represents the experiment results of a chosen category that

Company		Category 1 %	Category 2 %	Category 3 %
CTC	Category 1	100	9	8
	Category 2	41	100	22
	Category 3	28	16	100
FPC	Category 1	100	15	10
	Category 2	34	100	13
	Category 3	17	9	100
TSMC	Category 1	100	17	6
	Category 2	60	100	16
	Category 3	27	19	100
UMC	Category 1	100	16	3
	Category 2	64	100	8
	Category 3	25	16	100
ACER	Category 1	100	25	8
	Category 2	12	100	6
	Category 3	18	28	100
ASUS	Category 1	100	27	3
	Category 2	43	100	7
	Category 3	21	29	100

Table V.
Correlation among categories for the search term of each company

is the category of interest when the precision rates were calculated. The first column is the precision rate of the experiment without the ontology- based query expansion. This would be the baseline. The second column is the average precision rate for all test cases of the experiment with the ontology-based query expansion. The number of all test cases is 1,331, resulting from 11 multiplied by 11 multiplied by 11, since the weight of each chosen ontology structure increased by 10 from 0 to 100, and there are three categories chosen in the experiment. The third column is the average precision rate of the best cases of the experiment with the ontology-based query expansion. The best cases are cases in which the weight of the category of interest is higher than the weights of the other two categories. Among 1,331 test cases, there are a total of 385 best cases in the experiment. The values of the third column are considered to be the upper bound of the performance of the system since they represent the highest precision rates in the system. The fourth column is the average precision rate of the worst cases of the experiment with the ontology-based query expansion. The worst cases are cases in which the weight of the category of interest is lower than the weights of the other two categories. Among 1,331 test cases, there are a total of 385 worst cases in the experiment. The values of the fourth column are regarded as the lower bound of the performance of the system since they represent the lowest precision rates in the system (see Table VI).

Conclusion

The experiment results show that with the aid of ontology the average precision rate of all cases, in general, is dramatically higher than the precision rate for the default search result. If a user's preference for ontology structures is accurately learned through the feedback loop, then the average precision rate is at least twice or even three times higher the precision rate for the default search result as demonstrated in the best cases, in which the weight of the county or category of interest is higher than the weights of the others.

The improvement of the precision rate can be attributed to three factors: the first factor is the intrinsically semantic relationship among the search keywords and the expanded keywords stored in the ontology structures. The second factor is the Lucene's proximity search expression, which restricts Lucene to finding words that are within a specified distance, resulting in highly relevant search results. The last factor shows the weights of the ontology structures. The striking difference between the best cases and the worst cases shows that weights of the ontology structures play an essential role in improving the precision rate.

Aside from the weighting of ontology structures for the search terms, the weight can be further extended to take personal preference into account, which means for each user, there is a hit counter for each ontology structure. In doing so, the weight of expanded terms can be decided based on both the pooled preference – the per search term hit count – and the personal preference – the per person hit count – of ontology structures. People who search for the same words, with different preferences for ontology structures, will see a different search result, which may be regarded as a result of personalization.

As the average query length has been steadily increased in the recent years from 1 word to 3.3 words, the keyword expansion technique becomes less effective in general though still helpful for the majority of web users. Contrary to automatic feedback

Table VI.
The summary of the experiment results for the search term of each company

Company	Category	Precision rate without ontology %	Precision rate with ontology in worst cases %	Precision rate with ontology in all cases %	Precision rate with ontology in best cases %	Highest achievable precision rate %
CTC	1	50	78	92	100	100
	2	20	41	53	68	100
	3	15	18	33	48	100
FPC	1	40	66	83	95	100
	2	35	37	49	67	100
	3	20	27	48	72	100
TSMC	1	65	72	89	100	100
	2	30	64	73	88	100
	3	0	20	40	67	100
UMC	1	40	65	85	100	100
	2	50	60	74	92	100
	3	10	0	17	42	100
ACER	1	55	44	66	87	100
	2	50	40	64	86	100
	3	10	18	27	40	100
ASUS	1	80	82	93	100	100
	2	70	71	81	92	100
	3	10	20	24	32	100

learning methods like the one we propose in this paper, there are alternative approaches to improve the accuracy of query expansion. Google, for example, has conducted experiments to utilize the clustering technology (Hicks, 2004; Search Engine Lowdown, 2004), trying to make Google better at understanding the multiple meanings of a word, and its relationship to other words. Other companies such as Endeca and Exalead have recently provided innovative interfaces to allow users actively refine their initial search queries, which may result in better search results for experienced users. This continuing momentum of innovation in the field of information retrieval clearly demonstrates its importance and influence on our daily life.

References

Bogatin, D. (2006), "Yahoo: searches more sophisticated and specific", ZDNet, available at: http://blogs.zdnet.com/micro-markets/index.php?p = 27

Brin, S. and Page, L. (1998), "The anatomy of a large-scale hypertextual web search engine", *Computer Networks and ISDN Systems*, Vol. 33, pp. 107-17.

Enquiro Search Solutions Inc. (2005), "Did-it, Enquiro, and Eyetools uncover Search's Golden Triangle", available at: www.enquiro.com/eye-tracking-pr.asp

Harman, D. (1988), "Towards interactive query expansion", *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 321-31.

Hicks, M. (2004), "Google sets slights on clustering, translation", *eWeek*, available at: www.eweek.com/article2/0,1759,1668357,00.asp

Jarvelin, K., Kristensen, J., Niemi, T., Sormunen, E. and Keskustalo, H. (1996), "A deductive data model for query expansion", *Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 235-49.

Jing, Y. and Croft, W.B. (1994), "An association thesaurus for information retrieval", *Proceedings of RIAO*, pp. 146-60.

Jones, S., Gatford, M., Robertson, S., Hancock-Beaulieu, M. and Seeker, J. (1995), "Interactive thesaurus navigation: intelligence rules OK?", *Journal of the American Society for Information Science*, Vol. 46 No. 1, pp. 52-9.

Kristensen, J. (1993), "Expanding end-users' query statements for free text searching with a search-aid thesaurus", *Information Processing & Management*, Vol. 29 No. 6, pp. 733-44.

Lu, X.A. and Keefer, R.B. (1995), "Query expansion/reduction and its impact on retrieval effectiveness", *The Third Text Retrieval Conference (TREC-3)*, pp. 231-9.

Pu, H.T. and Yang, C. (2003), "Enriching user-oriented class associations for library classification schemes", *The Electronic Library*, Vol. 21 No. 2, pp. 130-41.

Rocchio, J. (1971), "Relevance feedback in information retrieval", in Salton, G. (Ed.), *The SMART Retrieval System: Experiments in Automatic Document Processing*, Prentice-Hall, Upper Saddle River, NJ.

Salton, G. and Buckley, C. (1990), "Improving retrieval performance by relevance feedback", *Journal of the American Society for Information Science*, Vol. 41, pp. 288-97.

Search Engine Inc. (2003), "Search engine partnerships", available at: www.searchengines.com/partnerships.html.

Search Engine Lowdown (2004), "Web 2.0 – exclusive demonstration of clustering from Google", available at: www.searchenginelowdown.com/2004/10/web-20-exclusive-demonstration-of.html

Voorhees, E. (1994), "Query expansion using lexical-semantic relations", *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 61-9.

Yang, C., Chen, L.C. and Peng, C.Y. (2006), "Developing and evaluating an IT specification extraction system", *The Electronic Library*, Vol. 24 No. 6, pp. 832-46.

Further reading

Hotchkiss, G., Garrison, M. and Jensen, S. (2004), "Search engine usage in North America", available at: www.enquiro.com/research.asp

Xu, J. and Croft, W. (1996), "Query expansion using local and global document analysis", *Proceedings of the 19th Annual International ACM SIGIR Conference*, pp. 4-11.

About the authors

Chyan Yang received his PhD in Computer Science from the University of Washington, Seattle. He also holds an M. in Information Science from Georgia Institute of Technology, a MBA in Management Science from National Chiao Tung University and received his BS in EE from National Chiao Tung University. Between 1987 and 1992 he worked as an assistant professor in the Department of Electrical and Computer Engineering at the US Naval Postgraduate School at Monterey, California. From 1992-1995 he was with the Institute of Management Science, National Chiao Tung University, Taiwan as an associate professor. Chyan Yang is now a Professor and Chairman in the Institute of Business and Management, and Director of EMBA program at National Chiao Tung University, Taiwan. He has been an IEEE senior member since 1992 and has worked as an advisor to several IT companies. With more than 60 journal papers and 90 conference papers published, his current research interests include information management, and strategic management. He is the corresponding author and can be contacted at: professor.yang@gmail.com

Keng-Chieh Yang is a PhD student of the Institute of Information Management, National Chiao Tung University (NCTU), Taiwan. Before he joined NCTU, he served as an associate researcher in Mechanical Industry Research Laboratories, Industrial Technology Research Institute. He has also been a senior engineer in Macronix International Co. Ltd and a system analyst in the Fulbright Foundation, Taiwan.

Hsu-Chieh Yuan is an engineer in a high tech company in Taiwan. He holds an MBA from the Institute of Information Management, National Chiao Tung University (NCTU), Taiwan.