LSI Based Relevance Computation for Topical Web Crawler

Gurmeen Minhas
UIET, Panjab University, Chandigarh, India
Email: gurmeenminhas@gmail.com

Mukesh Kumar
UIET, Panjab University, Chandigarh, India
Email: mukesh_rai9@yahoo.com

Abstract—**Today, size of the web is exceptionally large. And this size is increasing rapidly. Huge number of web pages and web sites are being added each day. Hence, results which are effective, factual and authentic are needed. A simple crawler cannot cover each web page as it would take polynomial time to do so. In order to overcome such issues, this paper proposes an algorithm to develop an efficient, focused, domain specific crawler using LSI (Latent Semantic Indexing). This algorithm makes the crawler highly efficient in downloading relevant documents, thus, avoiding over-heads and resource wastage, and also increases the precision and recall values of the IR system developed on it.**

Index Terms— **Crawling, focused crawler, latent semantic indexing, domain specific crawler.**

I. Introduction

The World Wide Web Worm (WWWW) is one of the first web engines. According to a survey in 1994, WWWW had an index of 110,000 web pages and web documents [9]. However creating a search engine which fulfils the present day web requirements is a challenging task. Today we need a fast crawling technology to gather the web documents and keep them up to date. A web search engine is designed to search for information on the World Wide Web [9]. A search engine basically has three steps, crawling the web, indexing and searching.

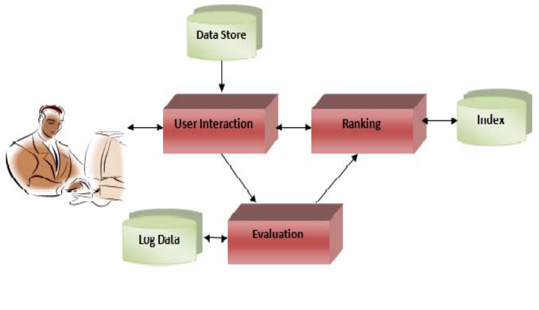

Figure 1. General Searching Process

Crawling the web means to download the documents present on the web which can be used for later queries. This task is performed by a web crawler. A software or a computer program that browses the World Wide web in a procedural, mechanized manner or in an orderly form is known as a web crawler. Other terms for web crawlers are ants, automatic indexers, bots , web spiders, web robots. The process is called web crawling or spidering [4]. Search engines use crawling as a part of its process to store and provide up-to-date data. Main work of a web crawler is to create a copy of all the pages it visits, for later processing by the search engine. Index is created in search engines so that the data is organized and quick results can be given to the user for their queries. Indexes are built based on the number of instances and position of particular words and then efficient ranking is implemented. The ranking of web pages are usually based on various factors such as number of times a word is being used in a document or the semantic structure of the content etc. Some ranking algorithms form the basis to calculate the score of the documents. The documents are ranked so that more relevant results are returned to the user in response to the user's query. The query input is taken from the user through the user interface of a search engine.

Focused search engines are domain-specific search engines which reduces the search margin which somehow increases the search accuracy. A focused search engine has a focused crawler at its heart, which gathers and updates information from the web [34]. A focused crawler is also known as topical crawler. Topical crawlers move over all web pages which are related to a particular subject, beginning from some relevant seed pages. The topical crawlers while travelling the web will analyze each hyperlink and try to figure out which link may be relevant to the subject. The relevant links are chosen and the irrelevant ones are abandoned [34]. Therefore, a focused crawler is the one which attempts to download only those web pages which are relevant to a pre-defined topic. A Focused Crawler is described as a mechanism which seeks, acquires, indexes, and maintains pages on a specific set of topics that represent a relatively narrow segment of the Web [6].

II. Related Work

doi:10.4304/jetwi.5.4.401-406

Due to the broad size of the web and the general crawling and indexing mechanism, results achieved are of low precision. As a result of this, the present day scenario demands specialized and focused crawlers. This section discusses some of the methods that have been used for the purpose of information retrieval and for constructing focused crawlers.

A. Backward and Forward Link Count

The idea that the number of times a paper is cited, has an impact on the importance of that particular paper, and this forms the basis for the concept of link counts [16]. So it is commonly regarded, that, a page that is linked by many other pages on the web will be more useful as compared to the page that is linked by a lesser number of other pages that is, a page which is referred scarcely is considered less important.

Suppose that we have a web page say P, and I(P) is the measure of importance of page P. So according to backward-link count metric, the importance I(P) of page (P), will be measured by the number of other pages on the web that have links pointing to page P, as shown in figure 2.

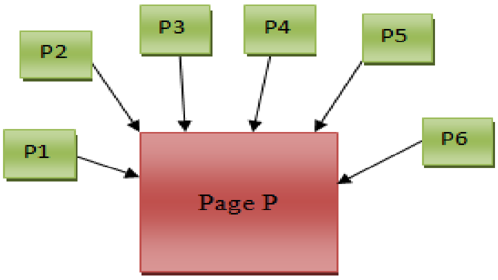

Figure 2. Backward-Link count

The other metric is the forward-link count shown in figure 3. According to this metric, a page that contains many outgoing links is treated important, since it may be a web directory or a web resource depository.

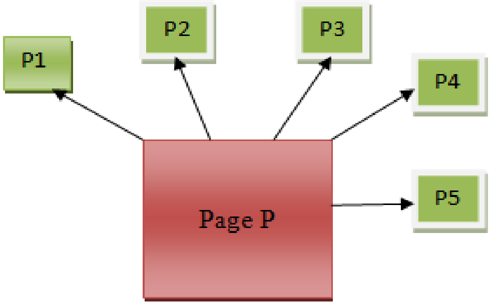

Figure 3. Forward-Link count

B. Page Rank

The original page rank metric was described by Sergey Brin and Lawrence Page [9]. To estimate the importance of a page, this metric makes use of the link structure of the page. Suppose
I(P)= page rank value of page P

I(Q)= page rank value of any page Q

B(P)= set of all pages that have a link pointing to P

c(Q)= number of all links in page Q

$I(P) = \sum_{Q \,\square\, B(P)} I(Q) / c(Q)$

Therefore, page rank of page P, that is, I(P), depends on the page rank value of page Q, that is, I(Q), where Q belongs to the set B(P), divided by number of links in page Q.

C. HITS(Hyperlink-Induced Topic Search)

Hyperlink-Induced Topic Search (HITS) is also known as hubs and authorities. HITS makes use of the link structure of the web, in order to discover and rank pages relevant for a particular topic. This algorithm was originally developed by Jon Kleinberg[10].

According to Jon Kleinberg [10] , the way a human user searches a document is very contrary and complicated than the process of just matching a list of query words with a list of documents. In this scheme, every web page is assigned two scores- hub score and authority score. So corresponding to a query two ranked lists are made [10]. The ranking of one list is based on hub scores and ranking of the other list is based on authority scores. Let us imagine that we have a query, "facts about formula 1 racing car ". The official page of formula 1 would be the authoritative source of information on the topic. Such pages will be called authorities. On the other hand there must be many web pages which would be containing lists of links to the authoritative web pages on a particular topic. These are known as hub pages. These hub pages are not sources of topic specific information but accumulation of links on that topic. A good hub page is the one that points to many good authorities, and a good authority page is one that is pointed to by many good hub pages. Suppose we have a set of web pages which are good hubs and authorities and these are having hyperlinks among them. The hub score and authority score for every web page is calculated. We have a web page v , hub score h(v) , authority score a(v) , v → y means a hyperlink exist from v to y. We have the following equations[1]

$$h(v) \leftarrow \sum_{v \to y} a(y)$$

$$a(v) \leftarrow \sum_{y \to v} h(y)$$

According to the first equation hub score of page v is equal to the sum of authority scores of the pages it links to. So if v links to pages with high authority scores then its hub score will be high. According to second equation if page v is linked to by good hubs then its authority score will be high.

D. Text Categorization

Text categorization is also known as text classification, or topic spotting. It is the task of automatically sorting a set of documents into categories or classes or topics from a predefined set.

Text categorization is a supervised learning task in which pre-defined category labels are assigned to new documents based on a training set of labelled documents [11]. Yaug and Liu in [11] have discussed five categorization models which are Support vector machines(SVM), K-NearestNeighbour classification, Linear Least Squares Fit (LLSF), Naive Bayes Classifier(NB), andNeural Network Techniques(NNet).

E. Boolean Queries

Among the traditional methods of information retrieval is Boolean Retrieval model. The Boolean retrieval method uses Boolean operators and is the most straightforward technique of retrieval. The basis of Boolean model is set theory and Boolean algebra. Documents are expressed by the terms extracted from documents and queries are expressed as Boolean expressions. This model consists of a query which is just as a set of words. The queries usually consists of AND, OR, NOT. This model is an exact-match retrieval model which means that the query should be clear-cut and a document either matches the query or it does not, that is, result is 0 or 1.

Joon et all[15] have proposed a ranking algorithm which is thesaurus based that measures the relevance of documents and return the top ranking documents. This algorithm called as E-relevance algorithm gives the similarity score between a query and document[15].

F. Vector Space Model

Vector Space Model represents text documents as vectors. It is different from Boolean retrieval. Boolean retrieval method assigns binary weights 0 or 1, whereas, vector space model assigns non-binary weights to terms in documents and queries. Depending on these weights, the degree of similarity or the inter-relationship between a query and document is found out. According to this model, a space is created, in which both documents as well as queries are represented as vectors. The dimension of the vector is equal to the number of unique terms present in the document. Weights are assigned to terms and this weight is usually based on the number of occurrences of a term in a document, and, this is known as term-frequency. The other weighing scheme mostly used is the tf-idf, where idf is inverse document frequency .

G. Ontology based retrieval

An ontology represents knowledge as a set of concepts and relationships between those concepts for a specific domain. Ontology is a semantic based retrieval technique which understand the meaning of the concept of the user query. Ontologies are arranged in a taxonomy of concepts. Ontology includes description of concepts and its properties. It also describes various features and aspects of the concept.

H. Latent Semantic Indexing

Generally when we retrieve information it is based on exact matching, that is, the terms in the query are matched to those in the document. But sometimes we have certain documents which are relevant to the query but does not contain the exact words as present in the query. So in such cases it is advisable to use a mechanism that helps us to retrieve documents on the basis of conceptual meaning of the query and document. For this we use the concept of Latent Semantic Indexing. Latent Semantic Indexing is also known as Latent Semantic analysis. LSI is a technique that enables us to analyse relationships between terms and concepts occurring in a text. LSI uses a mathematical technique called Singular Value Decomposition (SVD). The main element of LSI is its ability to extract the conceptual content of text by building associations between the terms that have similar contexts. This technique is so called because it has the ability to relate terms that are semantically similar in some text. It uncovers the latent semantic structure of words in a text corpus. When a query is issued on a set of documents on which LSI has been applied, the results that we get will be the ones which are conceptually similar to the query even if the results do not contain same specific words. Latent Semantic Indexing starts with a term by document matrix. Then, Singular Value Decomposition (SVD) is used to decompose the term by document matrix into three matrices: T, a term by dimension matrix, S, a singular value matrix (dimension by dimension), and D, a document by dimension matrix. The number of dimensions is r, that is the rank of the term by document matrix. The original matrix can be obtained, through matrix multiplication of TSDt. In an LSI system, the T, S and D matrices are truncated to k dimensions [19].

III. Implementation and Graphs

The objective of our work includes the development of a term corpus specific to CAD (Computer Aided Design) domain. After which a crawling algorithm is developed which works on the scoring system based on LSI(Latent Semantic Indexing). Finally we evaluate performance by comparing anchor and document scores relevant for crawling to simple breadth-first algorithm (algorithm 1) and keyword based approach (algorithm 2). Our focused crawler builds its corpus , which is specific to CAD domain. Therefore, it is a model that works on the principle of selecting only those web documents, from whom as per algorithm 3 (LSI) , it can gain information with respect to CAD domain only. In this process it is intuitively reducing the uncertainty about the category of a document item being selected for crawling provided by knowing the value of feature Y. Here item Y are the seed keywords or URLs or future hyperlinks or the titles.

Since the ultimate goal of algorithm 3 or our focused web crawler is to build a dataset that would provide a high information gain when used by a search engine or query engine , the selection of URLs and keywords is very important as it would lead to burning of less resources. We are taking the advantage of highly optimized anchor tags , and also taking the advantage of vector semantic model in algorithm 3. By doing the above process , we thus improve the recall and precision of our overall system.

It is apparent from the graph for recall analysis, Figure 4 that the recall value varies from 28% to 40.5%, which reflects the completeness or sensitivity of our algorithm 3. The recall value here means less number of crawl jobs that are false negative in nature , or in simple words , crawling less number of web documents that were selected erroneously or those web URLs which were supposed to be rejected but got selected in URL crawl priority queue.

However, it can also be seen from the precision graph, Figure 5 that value remains around 59.4% and 66.03% which is otherwise difficult to obtain had not the algorithm 3 been implemented , because normally if recall value increases (in our case it is moderate) the precision often decreases, as it gets harder to precise when the sample space increases. But , in our result we can see that precision remains moderate , that means around 60% crawls are true positive in nature , or in simple words, the web documents which were supposed to be in priority queue were correctly selected.

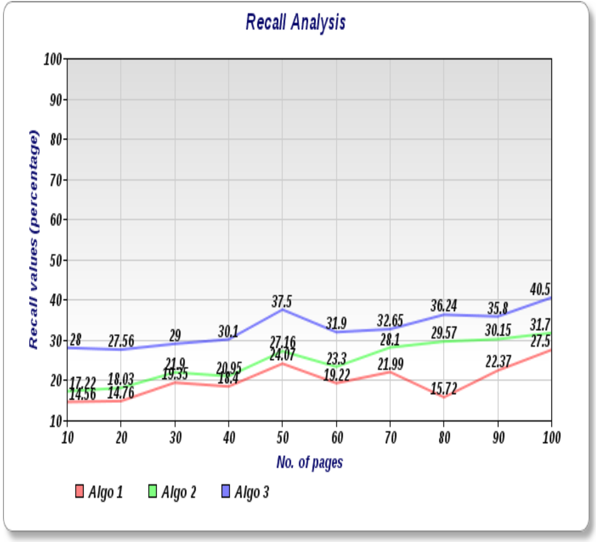

Figure 4. Recall Analysis Line Graph

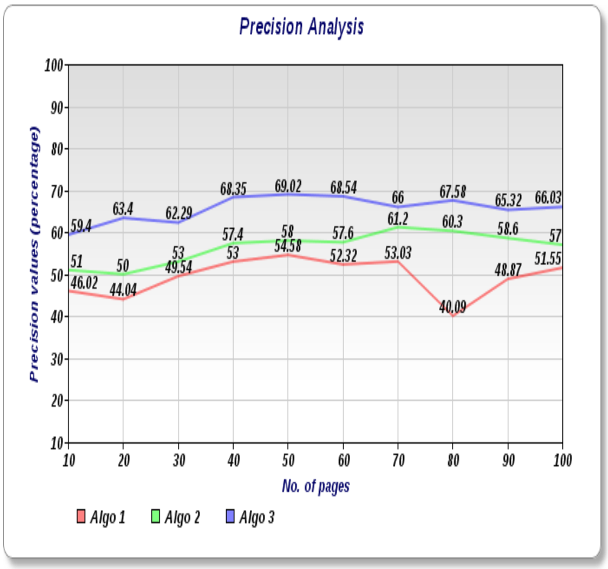

Figure 5. Precision Analysis Line Graph

IV. Conclusion

In this paper a domain specific focused crawler has been implemented. A domain specific crawler is useful for saving time and other resources since it is concerned with a particular domain. Hence we obtain highly relevant data which leads to high information gain and less resource wastage. Being in the field of engineering , computer aided design has been chosen as a domain to work on. Various methods of information retrieval have been studied and reviewed, and based on this literature survey, it was found that there is a requirement to build a crawler that takes into account the context of the words or phrases being searched for. LSI(Latent semantic Indexing) model is one such promising model in the field of information retrieval. LSI uses a mathematical technique known as Singular Value Decomposition. This model has the ability to extract the conceptual content of a body of text by looking for relationships between the terms of the text. The evaluation of the work has been done by using the recall and precision values. The values of my crawler has been compared to the recall and precision values of two other crawlers, which are, breadth first crawler and keyword based crawler. A breadth first crawler crawls the pages in the order they are encountered, without taking into account the relevancy and importance, as it continues in the direction wherever it finds the next link. A keyword based crawler makes use of the keywords supplied to the crawler. If the fetched pages contain 20% of these keywords, then that page is considered as relevant otherwise not. As apparent from the graph, precision values of a simple breadth first crawler is the least , then comes the keyword based crawler and finally the LSI based crawler. Hence it is clear that the performance of LSI based crawler is the most superior.

V. Future Scope

These days many information retrieval systems are being created based on taxonomies, ontologies, knowledge bases. The users want information based on particular domains which would help them save time and effort and would help them retrieve more relevant and useful results. However there is still lot to do in the field of domain specific web crawlers. Creation of more domain based crawlers in future is suggested in various areas such as chemistry , biology , medicine , etc. We can also add other machine learning algorithms like probabilistic algorithms , neural network etc which may result in even better precision.

REFERENCES

[1] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, "Introduction to Information Retrieval", Cambridge University Press. 2008.

[2] http://en.wikipedia.org/wiki/Information_retrieval.

[3] Dagobert Soergel "'Information Retrieval The scope of IR" , HCL Encyclopedia .

[4] http://en.wikipedia.org/wiki/Web_crawler.

[5] S.S. Dhenakaran and K. Thirugnana Sambanthan, " Web crawler - an overview", International Journal of Computer Science and Communication Vol. 2, No. 1, January-June 2011, pp. 265-267.

[6] Soumen Chakrabarti , Martin van den Berg , Byron Dom, " Focused crawling: a new approach to topic-specific Web resource discovery", Computer Networks 31 (1999) 1623–1640 (Elsevier).

[7] Qu Cheng, Xiamen Univ, Xiamen Wang Beizhan , Wei Pianpian, " Efficient focused crawling strategy using combination of link structure and content similarity ", IEEE International Symposium on IT in Medicine and Education, 2008.

[8] J. Cho, Hector Garcia Molina, Lawrence page, "Efficient Crawling through URL Ordering", paper presented at 7th international WWW Conference. April 1998. Brisbane, Australia.

[9] Sergey Brin , Lawrence Page, "The anatomy of a large-scale hypertextual Web search engine", Computer Networks and ISDN Systems 30 (1998) 107- 117 (Elsevier).

[10] J. Kleinberg, Authoritative sources in a hyperlinked environment, in: Proceedings of the Ninth Annual ACM-SIAM Symposium, Discrete Algorithms, January 1998, pp. 668-677.

[11] Yiming Yang, Xin Liu, " A re-examination of text categorization methods", Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval 1999.

[12] M,Deligenti, F Coetzel , S Lawrence , C Leegiles , M Gori , "Focused Crawling using Context Graphs" , presented in 26th International Conference on Very Large Databases,Cairo , Egypt,2000.

[13] Fan Wu, Ching-Chi Hsu, "Topic-specific crawling on the Web with the measurements of the relevancy context graph", Information Systems Volume 31 Issue 4-5, June, 2006. (Elsevier).

[14] D. Gibson, J. Kleinberg , "Inferring web communities from link topology" , Proceedings of the 9th ACM Conference on Hypertext and Hypermedia (HYPER-98), 1998, pp. 225–234.

[15] Joon ho lee, myoung ho kim, and yoon joon lee , "Ranking documents in thesaurus-based boolean retrieval systems", Information Processing and Management Vol. 30, No. 1. PP. 79-91. 1994.

[16] Wenlei Mao, Wesley W. Chu, " The phrase-based vector space model for automatic retrieval of free-text medical documents" , Data & Knowledge Engineering Volume 61 Issue 1, April, 2007 Pages 76-92 (Elsevier).

[17] Jibran Mustafa, Sharifullah Khan, Khalid Latif, "Ontology Based Semantic Information Retrieval", 4th International IEEE Conference on Intelligent Systems, 2008.

[18] Alex Thomo, "Latent Semantic analysis Tutorial" , http://www.engr.uvic.ca/~seng474/svd.pdf.

[19] April Kontostathis a and William M. Pottenger b, A Framework for Understanding Latent Semantic Indexing (LSI) Performance, International journal on Information Processing and Management, Volume 42, january 2006 (Elsevier).

[20] Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. , "Indexing by latent semantic analysis" in Journal of the American Society of Information Science,1999.

[21] Mohsen Jamali, Hassan Sayyadi, Babak Bagheri Hariri and Hassan Abolhassani, " A Method for Focused Crawling Using Combination of Link Structure and Content Similarity", Proceedings of the 2006 IEEE International Conference on Web Intelligence.

[22] Hongfei Yan Jianyong , Hongfei Yan , Jianyong Wang , Xiaoming Li , Lin Guo, "Architectural design and evaluation of an efficient Web-crawling system ", The Journal of Systems and Software 60 (2002) 185–193 (Elsevier).

[23] Anshika Pal, Deepak Singh Tomar, S.C. Shrivastava, " Effective Focused Crawling Based on Content and Link Structure Analysis", International Journal of Computer Science and Information Security, Vol. 2, No. 1, June 2009.

[24] Knut Magne Risvik *, Rolf Michelsen, " Search engines and Web dynamics", Computer Networks 39 (2002) 289–302 (Elsevier).

[25] Yajun Du ; Zhanshen Li , " The New Clustering Strategy and Algorithm Based on Latent Semantic Indexing", Natural Computation, 2008 . IEEE Fourth International Conference.

[26] A. Rungsawang, N. Angkawattanawit, " Learnable topic-specific web crawler", Journal of Network and Computer Applications 28 (2005) 97–114 (Elsevier).

[27] Zhumin Chen, Jun Ma, Jingsheng Lei, Bo Yuan, Li Lian , Ling Song, " A cross-language focused crawling algorithm based on multiple relevance prediction strategies", Computers and Mathematics with Applications 57 (2009) 1057_1072 (Elsevier).

[28] Robert C. Miller, Krishna Bharat, " SPHINX: a framework for creating personal, site-specific Web crawlers" Computer Networks and ISDN Systems 30 (1998) I I9- I30.

[29] Alexandros Batzios , Christos Dimou, Andreas L. Symeonidis, Pericles A. Mitkas, "BioCrawler: An intelligent crawler for the semantic web", Expert Systems with Applications 35 (2008) 524–530 (Elsevier).

[30] Cioara T. , Anghel I. , Salomie I. , Dinsoreanu M. , "A context-based Semantically Enhanced Information Retrieval Model", IEEE 5th International Conference on Intelligent Computer Communication and Processing, 2009.

[31] Aidan Hogan, Andreas Harth, Jürgen Umbrich, Sheila Kinsella, Axel Polleres, Stefan Decker ,"Searching and browsing Linked Data with SWSE: The Semantic Web search Engine", Web Semantics: Science, Services and Agents on the World Wide Web, Volume 9, Issue 4, December 2011 (Elsevier).

[32] G. Almpanidis, C. Kotropoulos, I. Pitas, "Combining text and link analysis for focused crawling—An application for vertical search engines" , Information Systems 32 (2007) 886–908 (Elsevier).

[33] Pooja Gupta , Ashok Sharma , J.P. Gupta, "A Novel Framework for Context Based Distributed Focused Crawler (CBDFC)", Int. J. Computer and Communication Technology, Vol. 1, No. 1, 2009.

[34] Hong-Wei Hao , Cui-Xia Mu, Xu-Cheng Yin , Shen Li, Zhi-Bin Wang, "An Improved Topic Relevance Algorithm for Focused Crawling", in IEEE International Conference on Systems, Man, and Cybernetics (SMC) , 2011.

[35] Hai Dong Hussain, F.K. Chang, E., "A survey in semantic web technologies-inspired focused crawlers", IEEE 3rd International Conference on Digital Information Management, 2008.

[36] Knut Magne Risvik, Rolf Michelsen, "Search engines and Web dynamics", International journal Computer Networks 39 (2002) 289–302 (Elsevier).

[37] Sotiris Batsakis, Euripides G.M. Petrakis, Evangelos Milios, "Improving the performance of focused web crawlers", Data & Knowledge Engineering 68 (2009) 1001–1013 (Elsevier).

[38] http://en.wikipedia.org/wiki/.NET_Framework.

[39] Todd A. Letsche , "Large-Scale Information Retrieval with Latent Semantic Indexing" ,

[40] Ritendra Datta , Dhiraj Joshi, Jia Li, James Z. Wang , " Image retrieval: Ideas, influences, and trends of the new age " , ACM Comput. Surv., Vol. 40, No. 2. , May 2008.

[41] http://www.creighton.edu/fileadmin/user/HSL/docs/ref/Searching_-_Recall_Precision.pdf