ISSN 2319 - 7595
Volume 2, No.2, March – April 2013
International Journal of Information Systems and Computer Sciences
Available Online at http://warse.org/pdfs/ijiscs02222013.pdf

Retrieval of Web Pages Using Integrated Content and Structured Exploration

Rekha Jain[1], Rupal Bhargava[2], Sulochana Nathawat[3], G.N Purohit[4]
[1, 2, 3, 4] Banasthali Vidyapith
[1]rekha_leo2003@yahoo.com
[2]bhargava.rupal@gmail.com
[3]nathawat.sulochana@gmail.com
[4]gn_purohitjaipur@yahoo.co.in

ABSTRACT

Web is the most unstructured and ever expanding repository. It has taken the living standards and access to information source to the peek. With the increase in demands of information retrieval it has become very important to analyze and study the web well to help users find the relevant data. One most common hindrance to efficient Information Retrieval is ambiguity in search terms entered by the user for searching data. In this paper we will be discussing different web mining and word sense disambiguation techniques. Also we are proposing an algorithm which will face such problems and give user the most relevant results.

Key words: Lexical Ambiguity, Web Mining, Word Sense Disambiguation

1. INTRODUCTION

Highlight In our day to day life, Internet has become a basic necessity. Complete world is accessible at one click. But as we know, with great advancement in technologies comes a great deal of challenges. We achieve milestones through our efforts yet strive to bear and tackle the after effects. Internet warehouse is vast and expanding rapidly. But the share of organized data is decreasing at almost the same rate. What needed is a way to logically and systematically use and maintain this datum. For this it is necessary to study the structure of web and web mining techniques. Another problem faced by the user is the ambiguity present in search terms, which is included knowingly or unknowingly by user. User always expects to get the relevant results in return of whatever he/she enters as search terms. There can be various type of ambiguities present in search terms. In this paper, we will be proposing an algorithm which would resolve lexical ambiguity of search terms and return the most relevant results to user.

2. WEB MINING

Web Mining is a process of extracting knowledge from web data where at least one structure or usage data is used in mining process [5]. The absolute process of extracting knowledge from web data is given in Figure 1 [7]. Patterns are recognized in raw data using mining tools and knowledge is then gained out of those pattern using different representation and visualization techniques.

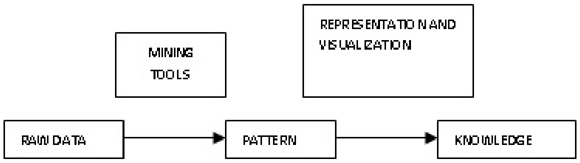


Figure 1: Process of Web Mining

3. WEB TAXANOMY

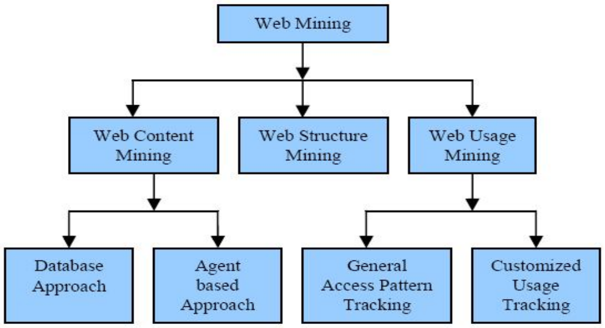


Figure 2: Web Mining Categories [1]

Web Mining is categorized on three basis, Web Content Mining, Web Structure Mining and Web Usage Mining. All the mining categories use different approaches to mine the data. Web Mining Categories are explained below in detail:

3.1. WEB CONTENT MINING

Web Content Mining is the process of capturing useful information from the contents of web documents. Content data corresponds to collection of data embedded in web pages to convey something to the user [5].

3.2. WEB STRUCTURE MINING

Word Structure Mining is extracting structural information of web [5]. Structure of web is considered as a graph that depicts web pages as nodes and hyperlinks as edges connecting two or more web pages.

3.3. WEB USAGE MINING

Web Usage Mining is the mining technique to discover interesting usage patterns from web data. Usage data captures identity or origing of web users along with their browsing behavior at website [5].

4. SENSE AMBIGUITY

When a word, phrase or a sentence has more than one interpretation it is said to be ambiguous. There are two type of ambiguity: Syntactic and Semantic.

4.1. SEMANTIC AMBIGUITY

When a word has more than one meaning it may result in semantic or lexical ambiguity. E.g. Shade can used to represent a color shade or it can be used as place to hide.

4.2. SYNTACTIC AMBIGUITY

When structure of a sentence leads to more than one possible interpretation. E.g. John saw the man on the mountain with a telescope. (Who has the telescope? John, the man on the mountain, or the mountain?)

We confront with ambiguities of natural language in daily use. However humans don’t find it much difficult to resolve such sense ambiguity. But for machine it is a great deal of work to resolve ambiguity. Hence for the purpose of resolving sense ambiguity we use Word Sense Disambiguation (WSD).

5. WORD SENSE DISAMBIGUATION & ITS APPROACHES

Word Sense Disambiguation is a task of identifying the correct meaning of a word in context at lexical level. It is a fundamental problem of Natural Language Processing. There are various distinguish approaches used for WSD [6]:

5.1. SUPERVISED APPROACH

This approach uses machine learning techniques for classifier to learn from labeled training sets, that is, sets of examples encoded in term number of features together with their appropriate sense label (or class) [6] [11].

5.2. SEMI SUPERVISED

In this approach both sense labeled and unlabeled data are employed in different proportions for a classifier to learn. This is most widely used approach due to the lack of training sets. There is only a partial amount of supervision [6] [11].

5.3. UNSUPERVISED APPROACH

This approach is based on unlabeled corpra, and do not exploit any manually sense tagged corpus to provide a sense choice for a word in context [6] [11].

5.4. KNOWLEDGE BASED (KNOWLEDGE RICH, DICTIONARY BASED)

It relies on the use of external lexical resources, such as machine readable dictionaries, thesauri, ontology, etc. [6] [11]

We have used Supervised Dictionary based approach in our proposed algorithm to disambiguate the search terms.

6. PROPOSED ALGORITHM

Proposed system disambiguates search terms to provide user with the efficient results. All users have to just go through a one-time registration which will help disambiguate search queries on his interest basis. This algorithm is designed to act as a layer onto Google search engine but it can be developed onto any other search engine. There are two phases in this algorithm: Pre Phase and Post Phase.

6.1. PRE-PHASE

It is requires to tokenize the search query and remove all the helping and stop words, so that no false negatives results are retrieved. After this, disambiguation is done on the basis of user interest in this phase.

Input: Search Query
Output: Modified Disambiguated Query

Algorithm

```
Step 1  Preprocessing of query string
          1.1  Tokenization
          1.2  Stemming
          1.3  Stop Word Removal
Step 2  If token contain polysemus word then
          Add sense or appropriate meaning on basis of user
          interest
```

6.2. MATCHING PHASE

In this phase disambiguated tokens of search string are compared with the database set of keyword, title, etc. to retrieve relevant results.

Input: Disambiguated set of tokens
Output: Results retrieved on basis of Page Rank algorithm

Algorithm

Step 1 Disambiguated set of tokens is passed to the search engine database where tokens are matched to find relevant results.
Step 2 Results retrieved are arranged according to value of Page Rank (Google's Ranking Mechanism)

6.3. POST-PHASE

This phase calculates the dynamic page rank using a count of exactly matched words and rearranges the results such that most appropriate result appears on the top.

Input: Result set retrieved from database of web pages
Output: Relevant Results

Algorithm

Step 1 Calculate Dynamic Page Rank based on the relevancy of retrieved result to the user search terms
Step 2 Sort the results according to the decreasing order of Dynamic Page Rank.

7. EVALUATION MEASURES

There are various number of evaluation measures available for evaluating the retrieved results. Some of them are:

7.1. PRECISION AND RECALL

Precision (P) is the fraction of retrieved documents that are relevant [3][8].

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(\text{relevant}|\text{retrieved}) \qquad (1)$$

Recall (R) is the fraction of relevant documents that are retrieved [3][8].

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} = P(\text{retrieved}|\text{relevant}) \qquad (2)$$

Table 1: Contingency Table

	Relevant	*Non Relevant*
Retrieved	*true positives (tp)*	*false positives (fp)*
Not retrieved	*false negatives (fn)*	*true negatives (tn)*

Where tp is relevant retrieved document
tn is non relevant documents which are not retrieved
fp is non relevant retrieved document
fn is relevant documents which are not retrieved

7.2. AVERAGE PRECISION

Average precision computes the average value of $p(r)$ over the interval from $r = 0$ to $r = 1$[4].

$$\text{AveP} = \int_0^1 p(r)dr. \qquad (3)$$

7.3. MEAN AVERAGE PRECISION

Mean average precision for a set of queries is the mean of the average precision scores for each query [4].

Where Q is the number of queries.

$$\text{MAP} = \frac{\sum_{q=1}^{Q} \text{AveP}(q)}{Q} \qquad (4)$$

7.4. RECIPROCAL RANK

Reciprocal Rank is a statistical measure for evaluating any process that produces a list of possible responses to a query, ordered by probability of correctness. Reciprocal rank is the inverse of the rank of first correct answer. [4]

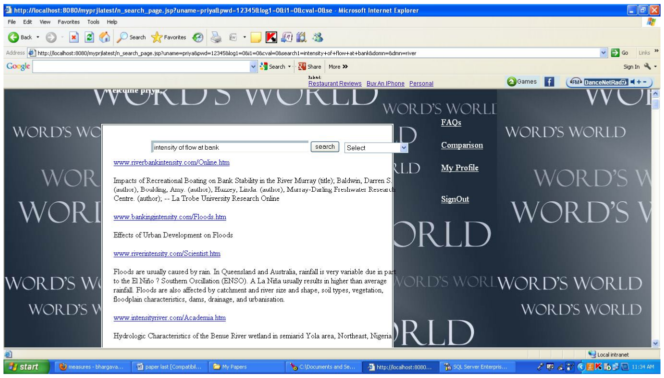
$$Reciprocal\ Rank = \frac{1}{rank} \qquad (5)$$

Mean Reciprocal Rank is the average of the reciprocal ranks of results for a sample of queries Q [4]:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \qquad (6)$$

8. EXPERIMENTAL RESULTS

Dynamic Page Rank algorithm resolves ambiguity of various words and phrases. Few examples of phrase based retrieval are:
Case1: we have taken a phrase intensity of flow at bank. It can have two meaning. One can refer to intensity of flow of water at river bank and another can be intensity of flow of money at bank which is a financial institution. Results for the same are shown in figure 3 and 4.



http://localhost:8080/myprjlatest/n_search_page.jsp?uname=priya&pwd=12345&log1=0&i1=0&cval=0&se - Microsoft Internet Explorer

Address http://localhost:8080/myprjlatest/n_search_page.jsp?uname=priya&pwd=12345&log1=0&i1=0&cval=0&search1=intensity+of+flow+at+bank&domn=&dmn=river

Welcome priya,

intensity of flow at bank

www.riverbankintensity.com/Online.htm

Impacts of Recreational Boating on Bank Stability in the River Murray (title); Baldwin, Darren S. (author), Boulding, Amy. (author), Huzzey, Linda. (author), Murray-Darling Freshwater Research Centre. (author); -- La Trobe University Research Online

www.bankingintensity.com/Floods.htm

Effects of Urban Development on Floods

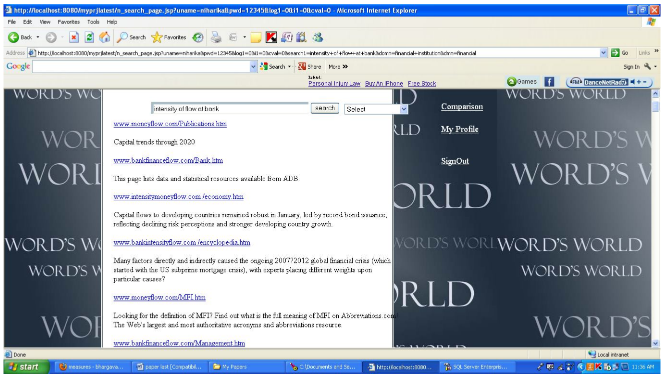
www.riverintensity.com/Scientist.htm

Floods are usually caused by rain. In Queensland and Australia, rainfall is very variable due in part to the El Niño ? Southern Oscillation (ENSO). A La Niña usually results in higher than average rainfall. Floods are also affected by catchment and river size and shape, soil types, vegetation, floodplain characteristics, dams, drainage, and urbanisation.

www.intensityriver.com/Academia.htm

Hydrologic Characteristics of the Benue River wetland in semiarid Yola area, Northeast, Nigeria

Figure 3: Intensity of flow at bank (river)


http://localhost:8080/myprjlatest/n_search_page.jsp?uname=niharika&pwd=12345&log1=0&i1=0&cval=0 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Go Links
Search Share More
Sign In
Personal Injury Law Buy An IPhone Free Stock
Games DanceNetRadio
search Select
Comparison
My Profile
SignOut
Done
Local intranet
start
11:36 AM

Address: http://localhost:8080/myprjlatest/n_search_page.jsp?uname=niharika&pwd=12345&log1=0&i1=0&cval=0&search1=intensity+of+flow+at+bank&domn=financial+institution&dmn=financial

intensity of flow at bank

www.moneyflow.com/Publications.htm

Capital trends through 2020

www.bankfinanceflow.com/Bank.htm

This page lists data and statistical resources available from ADB.

www.intensitymoneyflow.com /economy.htm

Capital flows to developing countries remained robust in January, led by record bond issuance, reflecting declining risk perceptions and stronger developing country growth.

www.bankintensityflow.com /encyclopedia.htm

Many factors directly and indirectly caused the ongoing 2007?2012 global financial crisis (which started with the US subprime mortgage crisis), with experts placing different weights upon particular causes?

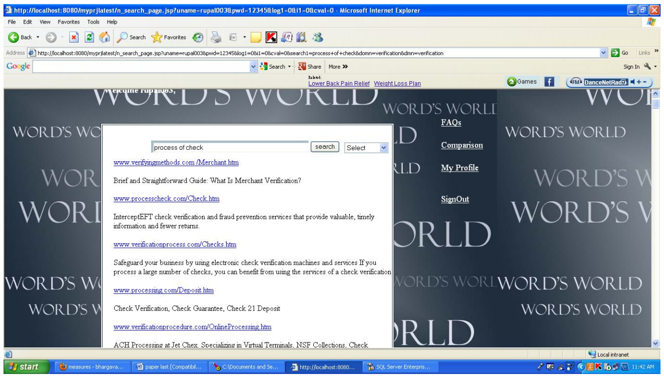
www.moneyflow.com/MFI.htm

Looking for the definition of MFI? Find out what is the full meaning of MFI on Abbreviations.com! The Web's largest and most authoritative acronyms and abbreviations resource.

www.bankfinanceflow.com/Management.htm

Figure 4: Intensity of flow at bank (financial institution)

Case2: We have taken another phrase process of check. This phrase can have two meanings. One can refer to process of check or verification and another can process of check and mate in chess. Results for the same are shown in figure 5 and 6.


http://localhost:8080/myprjlatest/n_search_page.jsp?uname=rupal003&pwd=12345&log1=0&i1=0&cval=0 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Go Links
Search Share More
Sign In
Lower Back Pain Relief Weight Loss Plan
Games DanceNetRadio
search Select
FAQs
Comparison
My Profile
SignOut
Local intranet
start
11:42 AM

Address: http://localhost:8080/myprjlatest/n_search_page.jsp?uname=rupal003&pwd=12345&log1=0&i1=0&cval=0&search1=process+of+check&domn=verification&dmn=verification

process of check

www.verifyingmethods.com /Merchant.htm

Brief and Straightforward Guide: What Is Merchant Verification?

www.processcheck.com/Check.htm

InterceptEFT check verification and fraud prevention services that provide valuable, timely information and fewer returns.

www.verificationprocess.com/Checks.htm

Safeguard your business by using electronic check verification machines and services If you process a large number of checks, you can benefit from using the services of a check verification

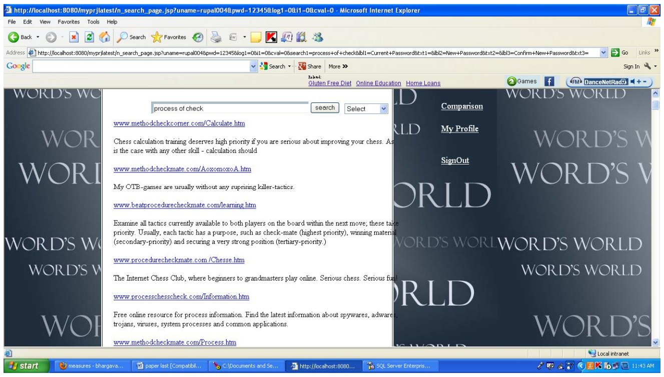
www.processing.com/Deposit.htm

Check Verification, Check Guarantee, Check 21 Deposit

www.verificationprocedure.com/OnlineProcessing.htm

ACH Processing at Jet Chex. Specializing in Virtual Terminals. NSF Collections. Check

Figure 5: Process of check (verification of something)


http://localhost:8080/myprjlatest/n_search_page.jsp?uname=rupal004&pwd=12345&log1=0&i1=0&cval=0 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Go Links
Search Share More
Sign In
Gluten Free Diet Online Education Home Loans
Games DanceNetRadio
search Select
Comparison
My Profile
SignOut
Local intranet
start
11:43 AM

Address: http://localhost:8080/myprjlatest/n_search_page.jsp?uname=rupal004&pwd=12345&log1=0&i1=0&cval=0&search1=process+of+check&lbl1=Current+Password&txt1=&lbl2=New+Password&txt2=&lbl3=Confirm+New+Password&txt3=

process of check

www.methodcheckcorner.com/Calculate.htm

Chess calculation training deserves high priority if you are serious about improving your chess. As is the case with any other skill - calculation should

www.methodcheckmate.com/AoxomoxoA.htm

My OTB-games are usually without any suprising killer-tactics.

www.beatprocedurecheckmate.com/learning.htm

Examine all tactics currently available to both players on the board within the next move; these take priority. Usually, each tactic has a purpose, such as check-mate (highest priority), winning material (secondary-priority) and securing a very strong position (tertiary-priority.)

www.procedurecheckmate.com /Chesse.htm

The Internet Chess Club, where beginners to grandmasters play online. Serious chess. Serious fun!

www.processchesscheck.com/Information.htm

Free online resource for process information. Find the latest information about spywares, adwares, trojans, viruses, system processes and common applications.

www.methodcheckmate.com/Process.htm

Figure 6: Process of check (chess)

Figure 7 depicts the comparison of average precision of page rank algorithm results and average precision of dynamic page rank algorithm results.

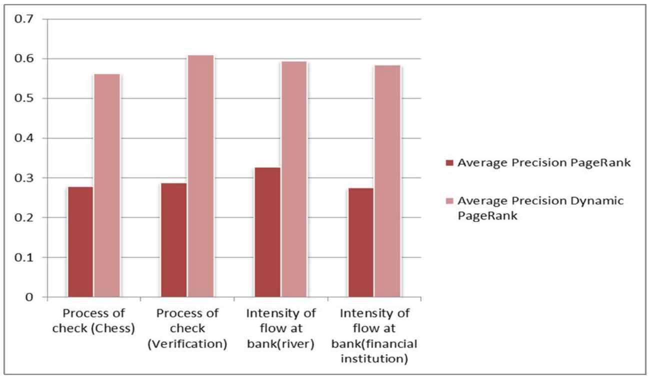


Figure 7: Comparative graph for Average Precision

Figure 8 depicts the comparison of reciprocal rank of page rank algorithm results and reciprocal rank of dynamic page rank algorithm results.

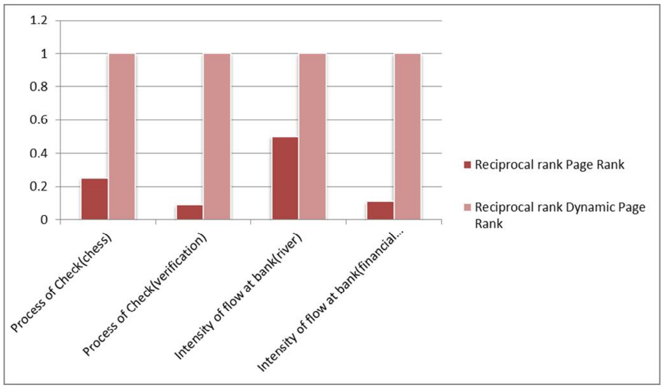


Figure 8: Comparative graph for Reciprocal Rank

Figure 9 depicts the comparison of mean average precision of page rank algorithm results and mean average precision of dynamic page rank algorithm results.

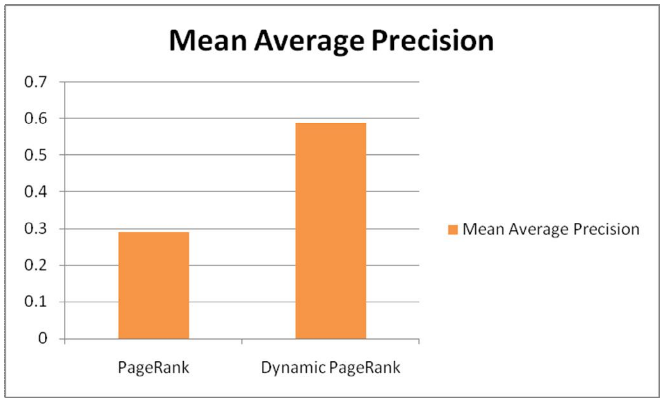


Figure 9: Comparative Graph of various Search queries for MAP

Figure 10 depicts the comparison of mean reciprocal rank of page rank algorithm results and mean reciprocal rank of dynamic page rank algorithm results.

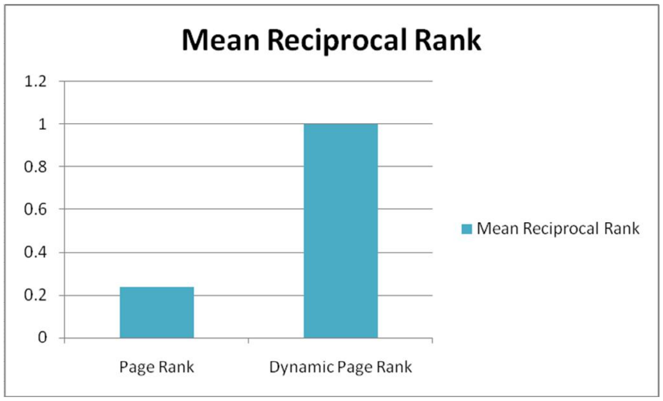


Figure 10: Comparative Graph of various Search queries for MRR

9. CONCLUSION

Web is ever expanding and with this expanding nature of web it is important to strengthen the retrieval methods in web mining. Our algorithm i.e. Dynamic Page Rank algorithm effectively disambiguated user search string and helped user to find most appropriate results. Few evaluation measures are also implemented to proof the relevancy and importance of work.

REFERENCES

1. Cooley, R., Mobasher, B., and Srivastava, J. "**Web mining: Information and pattern discovery on the World Wide Web**". In proceedings of the 9th IEEE International conference on Tools with Artificial Intelligence (ICTAI' 97), Newposrt Beach, CA, 1997.
2. Daniel Jurafsky and James H. Martin, **Speech and Language Processing**, Pearson Prentice Hall, 2009
3. **Evaluation in Information Retrieval**, Cambridge University Press. Feedback welcome, April 1, 2009
4. "**Information Retrieval**" available at http://en.wikipedia.org/wiki/Information_retrieval
5. Jaideep Srivastava , Prasanna Desikan , Vipin Kumar, **Web Mining – Accomplishments & Future Directions** available at "www.ieee.org.ar/downloads/Srivastava-tut-paper.pdf"
6. Navigli, R. 2009.**Word sense disambiguation: A survey**. ACM Comput. Surv. 41, 2, Article 10 (February 2009), 69 pages DOI = 10.1145/1459352.1459355 http://doi.acm.org/10.1145/1459352.1459355
7. Neelam Duhan, A.K Sharma and Komal Kumar Bhatia,"**Page Ranking Algoithms: A Survey**", In proceedings of the IEEE International Advanced Computing conference (IACC), 2009.
8. "**Precision and Recall**" available at http://en.wikipedia.org/wiki/Precision_and_recall
9. S. Brin, and L. Page, "**The Anatomy of a Large Scale Hypertextual Web Search Engine**", Computer Network and ISDN Systems, Vol. 30, Issue 1-7, pp. 107-117, 1998.
10. Schütze, H., "**Automatic word sense discrimination. Computational Linguistics**", 24(1): 97–123, 1998.
11. "**Word Sense Disambiguation**", available at http://en.wikipedia.org/wiki/Word-sense_disambiguation