ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 4, April 2013

Disambiguation of User Queries in Search Engines

Rekha Jain[*1], **Rupal Bhargava**[*2], **G.N Purohit**[*3]

***Department of Computer Science, Banasthali University, Jaipur, India**

Abstract —**With the increasing number of user queries to search engine. it has become necessary to study the area of Information Retrieval. User always expects to get the most appropriate and relevant results to the query. But when there is a ambiguity user usually do not get what he/she actually wants. This paper deals with ambiguity of the query and helps providing user with relevant results. Also to validate the work we have used evaluation measure Mean Average Precision.**

Keywords— Information Retrieval, Page Rank Algorithm, Precision, Average Precision, Mean Average Precision

I. Introduction

In recent years there have been a great number of user queries increased to the search engines. For all the queries humans are mostly relying on the search engines. Also user expects to get the most relevant and efficient results. There are many problems faced in doing so by the search engine, one of them is ambiguity of query. When the user enters a word beat, it has various meanings, beat can be related to boat, it can be in context of music, it can be in context of police etc. Such ambiguity in words leads to a depth study in area of Information Retrieval, NLP etc. In this paper we have discussed few aspects of these areas and proposed an algorithm to provide user more relevant results by disambiguating the queries of user. Also we have included some of the evaluation measures to proof the work.

II. Ranking Algorithms

Web Search engine uses different ranking algorithms to provide user with relevant results. Ranking Algorithms are generally used by web search engine to provide user results according to its priority by ranking the web pages. Different ranking algorithms use different aspects of user query to provide user with satisfactory relevant results. They are based on Web Structure Mining, Web Usage Mining, or Web Content Mining. All the three mining fields are playing a great deal of work in area of Information Retrieval. Many Ranking algorithms have been proposed and implemented till now, such as Page Rank Algorithm used by Google[6][8], HITS(Hypertext Induced Topic Search)[4][5], WPR(Weighted Page Rank)[10], etc.

III. Word Sense Disambiguation

Word sense Disambiguation is a task of automatically assigning sense to a polysemous word in a appropriate context. It helps find the correct meaning or sense to the sentence or query which has multiple meanings. It is a very challenging and important area of NLP which helps improving the performance of information retrieval and information extraction etc. When searching results, it is important to eliminate the senses of query or word which are not important to the user. Word Sense Disambiguation is the process that provide the direction to solve queries [9][11].

IV. Word Sense Disambiguation Approaches

There are two main approaches to WSD [11], Deep Approach that tries to access world of knowledge beforehand. Other approach is Shallow Approach that doesn't understand the text instead uses surrounding words to assign the meaning to text. Also there are four conventional approaches to WSD:

A. *Dictionary and Knowledge based method*

This method relies on dictionaries, thesauri, and knowledge bases they do not form a corpus for training. It assumes that words that are used together in text are related and the relation can be taken out from the meaning of words and sense of usage [11].

B. *Supervised*

This method assumes that context of text is enough to disambiguate the data. It uses sense annotated corpus to train the data [11].

C. *Semi-Supervised or Minimally Supervised*

This method is used by many Word sense disambiguation algorithms because of lack of training sets. It uses both annotated and un annotated data. A small set of annotated data is used to train initial data set. Then the initial classified data is used to train larger un annotated data. This process is repeated until whole corpus is consumed [9][11].

D. Unsupervised

It completely uses un annotated data. This method is biggest challenge for WSD researchers. It assumes that senses similar senses occur in similar contexts and thus senses can be induced from text by clustering word occurrences using some measure of similarity of context, a task referred to as word sense induction or discrimination.[9]

V. PROPOSED WORK

Our proposed work acts as a layer onto the Google's Page Rank algorithm but it can always use other ranking algorithms also. In our proposed work we have disambiguated the user query and calculated dynamic page rank for the results to rearrange the results according to user preferences.

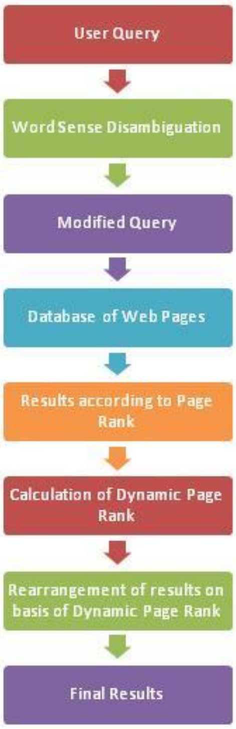

User Query
Word Sense Disambiguation
Modified Query
Database of Web Pages
Results according to Page Rank
Calculation of Dynamic Page Rank
Rearrangement of results on basis of Dynamic Page Rank
Final Results

Figure gorithm

VI. EVALUATION MEASURES

There are different measures available for evaluating performance of Information Retrieval. Some of them are:

A. Precision

Precision is the ratio of relevant retrieved documents to the retrieved documents. [3][7]

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \qquad (1)$$

B. Recall

Recall is the ratio of relevant retrieved documents to the relevant documents. [3][7]

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|} \qquad (2)$$

C. Average Precision

Average precision computes the average value of $p(r)$ over the interval from $r = 0$ to $r = 1$[3].

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

D. Mean average Precision

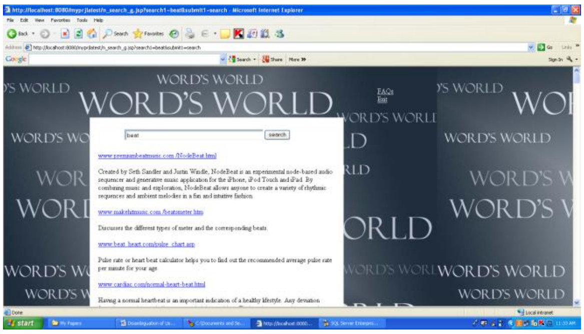
Mean average precision for a set of queries is the mean of the average precision scores for each query [3].

Where Q is the number of queries.

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

VII. EXPERIMENTAL RESULTS

There are various ambiguous words; we have chosen Beat which has various senses in context of heart, boat and music. The result of the simulation on the basis of Page rank algorithm and our Dynamic Page Rank Algorithm are shown in figure2, 3, 4,5. Also comparative graphs of Page Rank and Dynamic Page Rank for values of Average Precision and Mean Average Precision are given in figure 6,7.



beat

www.premiumbeatmusic.com /NodeBeat.html

Created by Seth Sandler and Justin Windle, NodeBeat is an experimental node-based audio sequencer and generative music application for the iPhone, iPod Touch and iPad. By combining music and exploration, NodeBeat allows anyone to create a variety of rhythmic sequences and ambient melodies in a fun and intuitive fashion.

www.makehitmusic.com /beatsmeter.htm

Discusses the different types of meter and the corresponding beats.

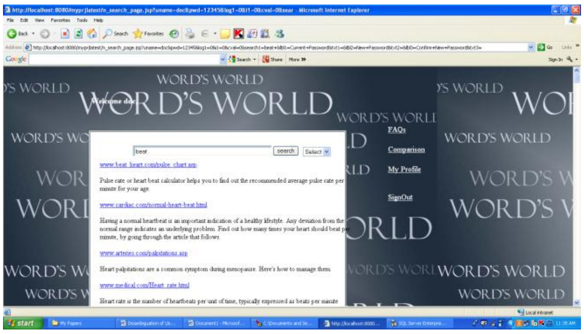
www.beat_heart.com/pulse_chart.asp

Pulse rate or heart beat calculator helps you to find out the recommended average pulse rate per minute for your age.

www.cardiac.com/normal-heart-beat.html

Having a normal heartbeat is an important indication of a healthy lifestyle. Any deviation

Figure 2: Results according to Page Rank



beat

www.beat_heart.com/pulse_chart.asp

Pulse rate or heart beat calculator helps you to find out the recommended average pulse rate per minute for your age.

www.cardiac.com/normal-heart-beat.html

Having a normal heartbeat is an important indication of a healthy lifestyle. Any deviation from the normal range indicates an underlying problem. Find out how many times your heart should beat per minute, by going through the article that follows.

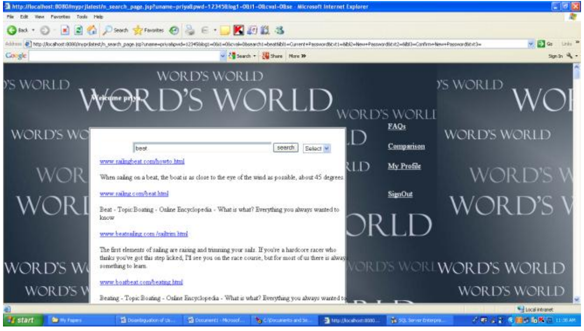
www.arteries.com/palpitations.asp

Heart palpitations are a common symptom during menopause. Here's how to manage them.

www.medical.com/Heart_rate.html

Heart rate is the number of heartbeats per unit of time, typically expressed as beats per minute

Figure 3: Results for beat (heart)



beat

www.sailingbeat.com/howto.html

When sailing on a beat, the boat is as close to the eye of the wind as possible, about 45 degrees.

www.sailing.com/beat.html

Beat - Topic:Boating - Online Encyclopedia - What is what? Everything you always wanted to know

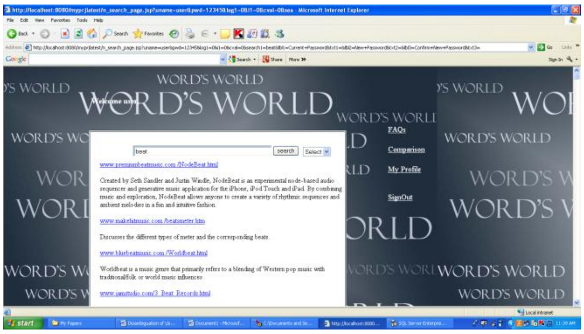
www.beatsailing.com /sailtrim.html

The first elements of sailing are raising and trimming your sails. If you're a hardcore racer who thinks you've got this step licked, I'll see you on the race course, but for most of us there is always something to learn.

www.boatbeat.com/beating.html

Beating - Topic:Boating - Online Encyclopedia - What is what? Everything you always wanted to

Figure 4: Results for beat (boat)


http://localhost:8080/myprjlatest/n_search_page.jsp?uname=user&pwd=123456&log1=0&l1=0&cval=0&sea - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Google Search Share More
Sign In
WORD'S WORLD
Welcome user
FAQs
Comparison
My Profile
SignOut
Local intranet
start

beat
search
Select

www.premiumbeatmusic.com /NodeBeat.html

Created by Seth Sandler and Justin Windle, NodeBeat is an experimental node-based audio sequencer and generative music application for the iPhone, iPod Touch and iPad. By combining music and exploration, NodeBeat allows anyone to create a variety of rhythmic sequences and ambient melodies in a fun and intuitive fashion.

www.makefatmusic.com /beatsmeter.htm

Discusses the different types of meter and the corresponding beats.

www.bluebeatmusic.com /Worldbeat.html

Worldbeat is a music genre that primarily refers to a blending of Western pop music with traditional/folk or world music influences.

www.jamstudio.com/3_Beat_Records.html

Figure 5: Results for beat (music)

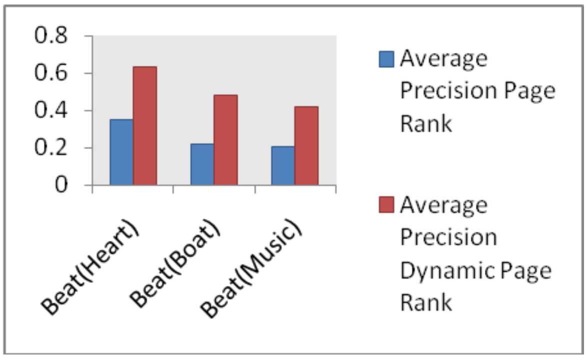


Figure 6: Comparative graph for Average Precision

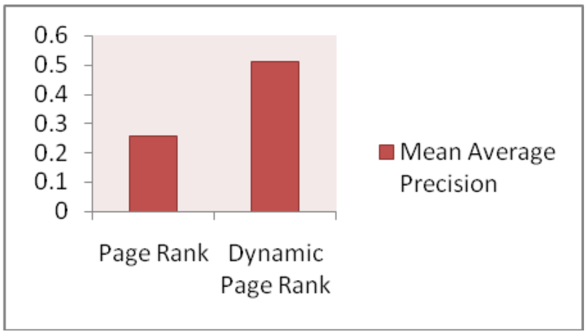


Figure 7: Comparative graph for Mean Average Precision

.

VIII. Conclusion

There are numerous ranking algorithms available for ranking web pages; all of them work on different perspectives. Page Rank algorithm is used by Google which is most popular search engine now days. But to solve the ambiguity faced by the user we have proposed a dynamic page rank algorithm which is acting as a layer on to the Page Rank algorithm and provides user much appropriate and relevant result. Also to validate the work we have included Mean Average Precision measure.

REFERENCES

[1] Ashutosh Kumar Singh, Ravi Kumar P, “A Comparative study of Page Ranking Algorithms for Information Retrieval”, International Journal of Electrical and Computer Engineering 4:7 2009

[2] Daniel Jurafsky and James H. Martin, Speech and Language Processing, Pearson Prentice Hall, 2009

[3] “Information Retrival” available at http://en.wikipedia.org/wiki/Information_retrieval

[4] J. Kleinberg, “Authoritative Sources in a Hyper-Linked Environment”, Journal of the ACM 46(5), pp. 604-632, 1999.

[5] J. Kleinberg, “Hubs, Authorities and Communities”, ACM Computing Surveys, 31(4), 1999.

[6] L. Page, S. Brin, R. Motwani, and T. Winograd, “The Pagerank Citation Ranking: Bringing order to the Web”. Technical Report, Stanford Digital Libraries SIDL-WP-1999-0120, 1999.

[7] “Precision and Recall” available at http://en.wikipedia.org/wiki/Precision_and_recall

[8] 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.

[9] Schütze, H., “Automatic word sense discrimination. Computational Linguistics”, 24(1): 97–123, 1998.

[10] W. Xing and Ali Ghorbani, “Weighted PageRank Algorithm”, Proc. Of the Second Annual Conference on Communication Networks and Services Research (CNSR ’04), IEEE, 2004.

[11] “Word Sense Disambiguation”, available at http://en.wikipedia.org/wiki/Word-sense_disambiguation

AUTHORS

Rekha Jain completed her Master Degree in Computer Science from Kurukshetra University in 2004. Now she is working as Assistant Professor in Department of “Apaji Institute of Mathematics & Applied Computer Technology” at Banasthali University, Rajasthan and pursuing Ph.D. under the supervision of Prof. G. N. Purohit. Her current research interest includes Web Mining, Semantic Web and Data Mining. She has various National and International publications and conferences.

Rupal Bhargava is pursuing her M.Tech in Computer Science from BanasthaliVidyapith, Rajasthan. She is undergoing the training of her M.Tech in supervision of Mrs. Rekha Jain. Her current research interest includes Web Mining, Semantic Web and Data Mining. She has published various papers in the conferences and journals.

Prof. G. N. Purohit is a Professor in Department of Mathematics & Statistics at Banasthali University (Rajasthan). Before joining Banasthali University, he was Professor and Head of the Department of Mathematics, University of Rajasthan, Jaipur. He had been Chief-editor of a research journal and regular reviewer of many journals. His present interest is in O.R., Discrete Mathematics and Communication networks. He has published round 40 research papers in various journals.