IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

ANALYZING AND PREDICTING USER'S BEHAVIOUR IN WEB SEARCH

Dr. Shubhashri Bose
Asstt. Prof, Deptt. Of Commerce
Vivekananda College
University of Delhi, India

Abstract : *Web Intelligence is an application of Business Intelligence software and methods that areused on Internet data. This paper provides the practical application of Information Retrieval using Web Intelligence. The Objective of this paper is to focus on the area of Web Mining that can easily synthesize to resolve the ambiguity between noun polysemous words. Here we have described the technique to resolve the ambiguities between polysemous words. Our previous work with the results has been described in the following section .The authors are trying to analyze the behavior of users and accordingto that they can predict the future expectation of user while performing web search.*

Index Terms*: *Web Intelligence, Information Retrieval, Web Mining, Polysemous words

INTRODUCTION

With the rapid growth of the Web, Information overload has become a serious concern as the explosive growth of resources available through the Internet. Web users are commonly overwhelmed by huge amount of information and are faced with the challenge of finding the most relevant and reliable information in a timely manner. For search engines, providing relevant pages of the highest quality to the users based on their queries becomes increasingly difficult and the level of difficulty increases with search for ambiguous words. This is because that some web pages are not self-descriptive and some links exist purely for navigational purposes and sometimes inappropriate association of meaning of the word which results in indifferent search from user's perspective. Two of the causes of ambiguity in natural languages are homonymy and polysemy, where homonymy refers to a word that has at least two entirely different meanings (such as "bark", which can mean the skin of a tree or the voice of a dog) while polysemy refers to a word which can take on two distinct, but related meanings (such as the "head" of the body, and the "head" of a department). The distinction between homonymy and polysemy is not always clear cut *(Akmajian, et al., 1990, Lyons, 1984, Kilgarriff, 1993)*. Lexical ambiguity covers both homonymy and polysemy. Here we are dealing with noun polysemous words. The Web offers new opportunities and challenges for many areas, such as business, commerce, marketing, finance, publishing, education, research and development. For computer scientists, the Web introduces many new research topics and provides a new platform to reconsider old problems. Web Intelligence is a new sub-discipline of computer science covering theories and technologies related to the Web. The scope of WI (Web Intelligence) as a research field, *(Zhong, et al., 2002)*, encompasses web information systems environments and foundations, ontological engineering, human-media interaction, web information management, web information retrieval, web agents, web mining and farming, and emerging web-based applications. It also aims at deepening the understanding of computational, logical, cognitive, physical, and social foundations as well as the enabling technologies for developing and applying Web-based intelligence and autonomous agents systems *(Liu, et al., 2003)*. We can study Web intelligence on at least four conceptual levels *(Zhong et al., 2002):*

- Network level – Internet-level communication, infrastructure, and security protocols, where intelligence comes from the Web adaptive to the user's surfing process.
- Interface level – Intelligent human-Internet interaction, e.g. personalized multimedia representation.
- Knowledge level – Representing (in machine-understandable formats) and processing the semantics of web data.
- Social level – Studying social interactions and behaviour of web users and finding user communities and interaction patterns.

Here in this paper we are dealing with one of major area of Web Intelligence (WI) that is Information Retrieval on the Web and along with it the focus is made on the interface level of WI i.e. to provide an interface which can disambiguate the noun polysemous words and hence results in relevant search from user's perspective.

RESEARCH METHODOLOGY

Web Mining

Web Mining is the technique used to crawl through various web resources to collect required information. It enables an individual to promote business, understanding marketing dynamics, new promotions floating on the internet etc. It is the use of data mining techniques to automatically discover and extract information from Web documents and services. Web mining should be decomposed into these subtasks *(R. Kosala,et.al, 2000*):

- Resource finding: It includes the task of retrieving intended Web documents.
- Information selection and pre-processing: It automatically selects and pre-processes specific information from retrieved Web resources. This step includes the transformation process of retrieved in Information Retrieval [IR] process from original data. These transformations cover
- removing stop words, finding phrases in the training corpus, transforming the representation to relational or first order logic form, etc.
- Generalization: It includes automatic discovery of the general patterns at individual Web sites as well as across multiple sites. Data mining techniques and machine learning are often used for generalization.
- Analysis: It refers to validation and/or interpretation of the mined patterns. In information and knowledge discovery process, people play very important role. This is important for validation and/or interpretation.

Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined.

Web Content Mining

Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. Text mining and its application to Web content has been most widely researched. Research activities in this field also involve using techniques from AI such as IR, Natural Language Processing [NLP], Image Processing and Computer Vision.

Web Structure Mining

The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages. Web Structure Mining can be regarded as the process of discovering structured information from the Web. This type of mining can be further divided into two categories based on the kind of structural data used. Hyperlinks: A Hyperlink is a structural unit that connects a Web page to different location, either within the same Web page or to a different Web page. A hyperlink that connects to a different part of the same page is called an Intra-Document Hyperlink, and a hyperlink that connects two different pages is called an Inter-Document Hyperlink. Document Structure: The content within a Web page can also be organized in a tree structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting Document Object Model [DOM] structures out of documents.

Web Usage Mining

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behaviour at a Web site. Capturing, Modelling and analyzing of behavioural patterns of users are the goal of this web mining category.

Word Sense Disambiguation

In the field of computational linguistics, the problem of lexical ambiguity is generally called Word Sense Disambiguation (WSD), and is defined as the problem of computationally determining which "sense" of a word is activated by the use of the word in a particular context. WSD is essentially a task of classification: where word senses are treated as classes, the context provides the evidence, and each occurrence of a word is assigned to one or more of its possible classes based on the evidence. WSD is an internal task in the NLP chain. It is used in many applications such as Machine Translation and Information Retrieval. Ambiguity has to be resolved in some queries. For instance, given the query "table" should return documents about furniture, round table or about the database table or the table tag. A similar problem arises for proper nouns such as *Raleigh* (bicycle, person, city, etc.). Current IR systems do not use explicit WSD, and rely on the user typing enough contexts in the query to only retrieve documents relevant to the intended sense (e.g. HTML "table" tag). Early experiments suggested that reliable IR would require at least 90% disambiguation accuracy for explicit WSD to be of benefit (*Sanderson, 1994*). More recently, WSD has been shown to improve cross-lingual IR and document classification (*Vossen, et al., 2006; Bloehdorn and Hotho, 2004;Clough and Stevenson, 2004*).

LITERATURE REVIEW

Ranking search results is a fundamental problem in Information Retrieval. Most common approaches primarily focus on similarity of query and a page, as well as the overall page quality (*R. Baeza-Yates,et.al,1999, S. Brin,et.al, 1997, G. Salton,et.al, 1983*). However, with increasing popularity of search engines, implicit feedback (i.e. the actions users take when interacting with the search engine) can be used to improve the rankings. Implicit relevance measures have been studied by several research groups. An overview of implicit measures is compiled by Kelly and Teevan (*Kelly,et.al, 2003*). This research, while developing valuable insights into implicit relevance measures, was not applied to improve the ranking of web search results in realistic settings.

In the middle of the 20th century, Word Sense Disambiguation in computational linguistics started emerging. The problem of WSD was first put forward in 1949 by Weaver who presented a mimeographed text discussing the need of WSD. He elaborated a very important problem related to the context used for disambiguating words: When disambiguating a certain word how many neighboring words should be taken into consideration. And then in the following decades researchers adopted many methods in an attempt to solve the problem of Automatic Word Sense Disambiguation, including: AI-based method, knowledge based method and corpus-based method (*Nancy Ide,et.al, 1998*). But the problem arises in this WSD technique because of unavailability of standardized system for word sense disambiguation, difficulty in obtaining large sense-tagged data sets adequately and along with it potential for WSD varies by task. (*Philip Resnik, et.al, 1997*).

PROPOSED WORK

The proposed system will make use of semantic knowledge in order to resolve ambiguity in entity extraction. The proposed technique identifies all possible meanings or senses of an entity and decides the most appropriate meaning of the entity inspired by the domain defined. The proposed model contains two phases. One is pre-processing and other is post-processing.

Framework

It elaborates the internal processing that has to be carried out for resolving ambiguity. Each unit of this framework is discussed in following sections.

Polysemous Word

Entity is defined here as a word that is categorized into the noun part-of-speech. Polysemous words are defined as "having or characterized by many meanings or could say the existence of several meanings for a single word or phrase". Here the polysemous word is obtained by the dissemination of the query string provided by the user for search. Knowledge about the word will be helpful in associating the most possible sense of an ambiguous entity.

Knowledge Repository

The repository will act as a store house which will provide all possible meanings of the word considered for search. This will be helpful in further processing of the system for optimizing the search result.

Extracted Sense List

Here the list of all meanings of the word is extracted and stored which would be supplied further information for the classification.

Sense Matching and Classification

The predefined domain would be compared with the extracted meanings of the incoming input. The relevant and most appropriate meaning of the word is retrieved and this would be provided to the search engine for further processing.

Optimized Search Result

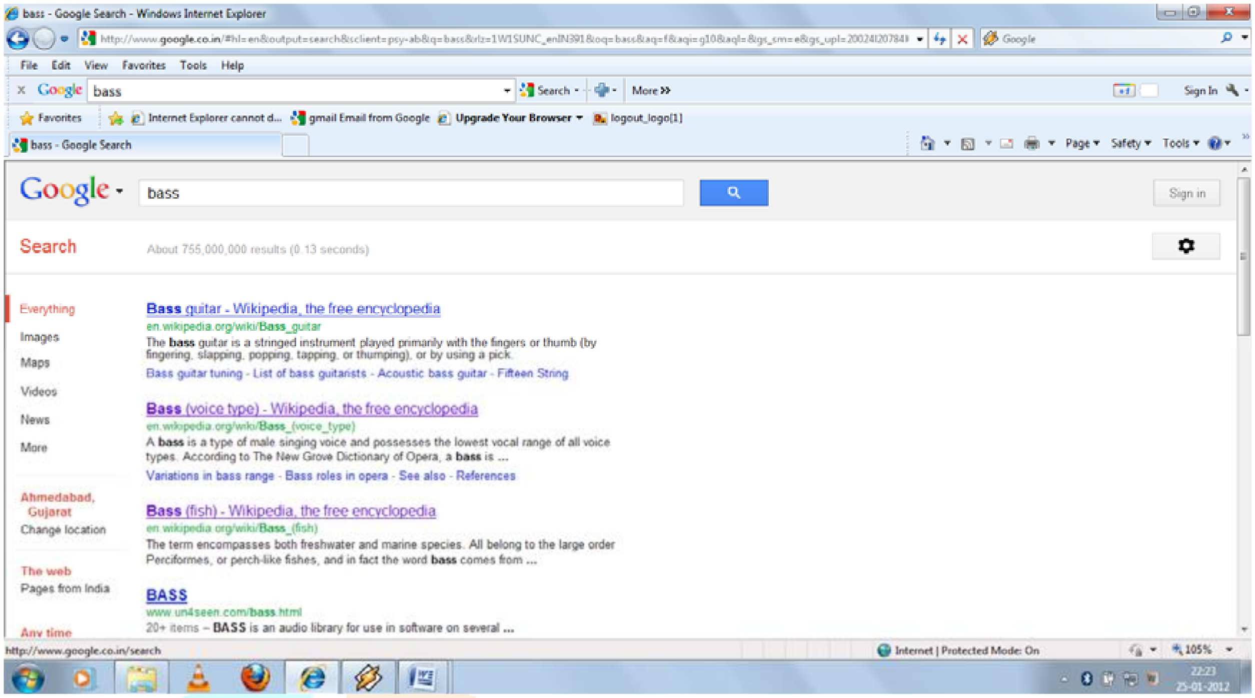
The result would be the list of relevant data searched by the user

Ambiguity Resolution

The resolution of ambiguity of a word is based on the predefined domain. This domain acts as the benchmark for evaluating and disambiguating the polysemous word. Effort is made to associate an appropriate meaning to the word in order to optimize the search result, which will influence and guide the further search. This will act as catalyst for producing optimized and relevant search result from user’s perspective.

Results

When the search made by musician or ichthyologist on search engines like Google for the ambiguous word say “bass”. The following sets of results are displayed.


bass - Google Search - Windows Internet Explorer
http://www.google.co.in/#hl=en&output=search&sclient=psy-ab&q=bass&rlz=1W1SUNC_enIN391&oq=bass&aq=f&aqi=g10&aql=&gs_sm=e&gs_upl=20024120784l
Google
File Edit View Favorites Tools Help
Google bass
Search
More »
Sign In
Favorites
Internet Explorer cannot d...
gmail Email from Google
Upgrade Your Browser
logout_logo[1]
bass - Google Search
Page Safety Tools
Google
bass
Sign in
Search
Everything
Images
Maps
Videos
News
More
Ahmedabad, Gujarat
Change location
The web
Pages from India
Any time
http://www.google.co.in/search
Internet | Protected Mode: On
105%
22:23
25-01-2012

About 755,000,000 results (0.13 seconds)

Bass guitar - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Bass**_guitar
The **bass** guitar is a stringed instrument played primarily with the fingers or thumb (by fingering, slapping, popping, tapping, or thumping), or by using a pick.
Bass guitar tuning - List of bass guitarists - Acoustic bass guitar - Fifteen String

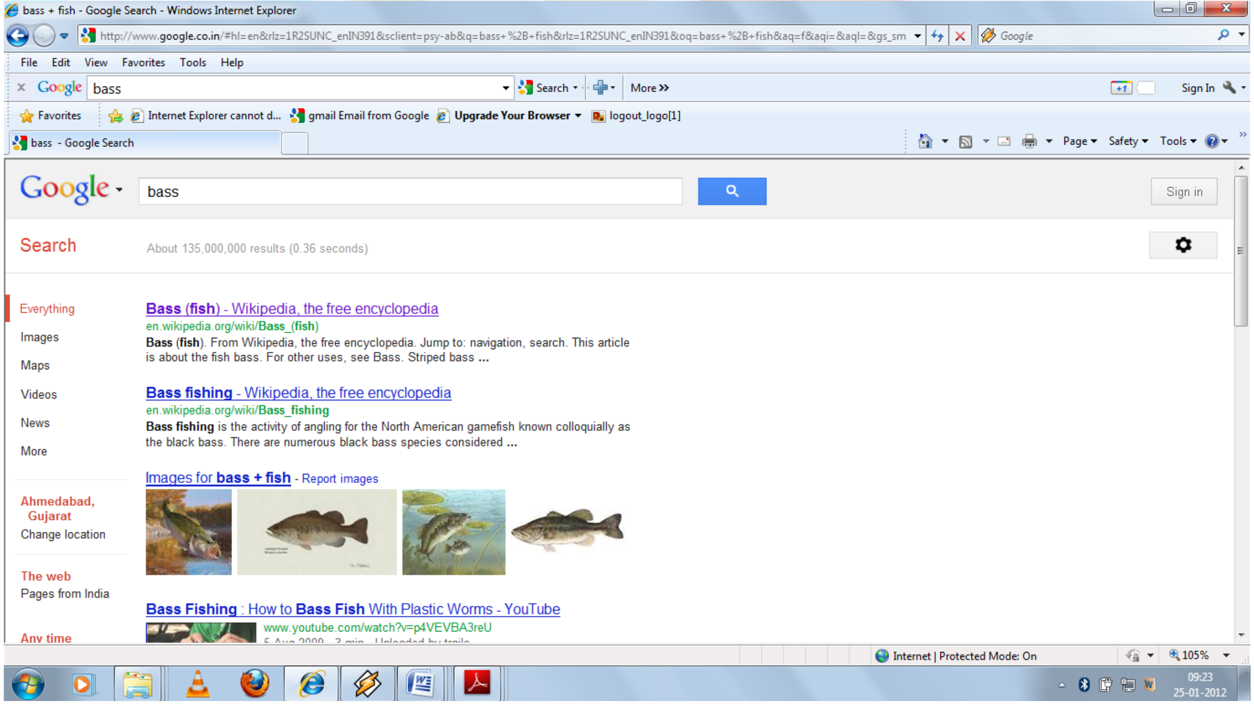
Bass (voice type) - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Bass**_(voice_type)
A **bass** is a type of male singing voice and possesses the lowest vocal range of all voice types. According to The New Grove Dictionary of Opera, a **bass** is ...
Variations in bass range - Bass roles in opera - See also - References

Bass (fish) - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Bass**_(fish)
The term encompasses both freshwater and marine species. All belong to the large order Perciformes, or perch-like fishes, and in fact the word **bass** comes from ...

BASS
www.un4seen.com/**bass**.html
20+ items – **BASS** is an audio library for use in software on several ...

Figure 1: Results displayed by Google

Figure 1 displays the interface of the proposed system. When the search is made by the user say Musician or Ichthyologist on the proposed system for the word say “Bass”, results will be shown as:


bass + fish - Google Search - Windows Internet Explorer
http://www.google.co.in/#hl=en&rlz=1R2SUNC_enIN391&sclient=psy-ab&q=bass+%2B+fish&rlz=1R2SUNC_enIN391&oq=bass+%2B+fish&aq=f&aqi=&aql=&gs_sm
Google
File Edit View Favorites Tools Help
Google bass
Search
More »
Sign In
Favorites
Internet Explorer cannot d...
gmail Email from Google
Upgrade Your Browser
logout_logo[1]
bass - Google Search
Page Safety Tools
Google
bass
Sign in
Search
Everything
Images
Maps
Videos
News
More
Ahmedabad, Gujarat
Change location
The web
Pages from India
Any time
Internet | Protected Mode: On
105%
09:23
25-01-2012

About 135,000,000 results (0.36 seconds)

Bass (**fish**) - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Bass**_(**fish**)
Bass (**fish**). From Wikipedia, the free encyclopedia. Jump to: navigation, search. This article is about the fish bass. For other uses, see Bass. Striped bass ...

Bass fishing - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Bass**_**fishing**
Bass fishing is the activity of angling for the North American gamefish known colloquially as the black bass. There are numerous black bass species considered ...

Images for **bass + fish** - Report images

Bass Fishing : How to **Bass Fish** With Plastic Worms - YouTube
www.youtube.com/watch?v=p4VEVBA3reU

Figure 2: Results displayed to Ichthyologist

CONCLUSION

As far now, we have identified the noun polysemous words and became successful in implementing the scenario for efficiently disambiguating the words from the user’s perspective. The scenario elaborates abetter way to disambiguate the dual meaning of the word during search and provides relevant information to the user. The technique works by first analyzing the word provided for search. if the word is ambiguous then it resolve its ambiguity by selecting the most appropriate meaning with respect to user’s perspective and then this precise sense will further guide the search for that word. The search result is generated by Google Search Engine as this technique forward the appropriate sense of ambiguous word to Google Search Engine resulting in an intelligent search. Hence the user can receive the pages that may be fruitful and relevant according to his/her needs. This forms the pre processing part of the proposed system.

Future Work

The future scenario includes the post processing technique of the proposed model which will generate more refined and intelligent search in terms of the contents relevant to the user's perspective. In post processing we would rearrange the pages retrieved by the query, so that the most of the relevant pages will occur at the top of the result

References

- Akmajian A., Demers R., Farmer A., & Harnish R., (1990). "Linguistics: An Introduction to Language and Communication", "Morphology: The Study of The structure of Words". (pp.11-52). Cambridge: MIT Press.
- Adam Kilgarriff., (1993). "Dictionary Word Sense Distinctions: An Enquiry into their Nature, Computers and the Humanities", 26 (1-2), pp 365-387.
- D. Kelly and J. Teevan., (2003). "Implicit Feedback for Inferring user Preference: A bibliography". In *SIGIR Forum.*
- G. Salton & M. McGill., (1983). Introduction to Modern Information Retrieval. McGraw-Hill.
- Liu J., Zhong N., Yao Y., & Ras Z.W., (2003). "The Wisdom Web: New Challenges for Web Intelligence (WI)". Journal of Intelligent Information Systems, 20(1), 5-9.
- Lyons J., (1984). "Language and Linguistics: An Introduction", Cambridge: Cambridge University Press.
- Nancy Ide and Jean Veronis., (1998). "Introduction to the special issue on word sense disambiguation: the state of the art". Computer Linguist. 24(1):2–40.
- Neha Singh and Rekha Jain., (2012). "Disambiguation Technique for Polysemous Word". International Journal of Research in Engineering & Applied Sciences, Volume 2, and Issue 2 (February 2012) ISSN: 2249-3905, pp 874-882.
- Philip Resnik, David Yarowsky., (1997). "A Perspective on Word Sense Disambiguation Methods and their Evaluation",in proceedings of SIGLEX '97, Washington, DC, pp. 79-86.
- R. Kosala, and H. Blockeel., (2000). "Web Mining Research: A Survey, SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining" Vol. 2, No. 1 pp 1-15.
- R. Baeza-Yates and B. Ribeiro-Neto., (1999). "Modern Information Retrieval", Addison-Wesley. S. Brin and L. Page., (1997). "The Anatomy of a Large-scale Hyper textual Web Search Engine", in *proceedings of WWW.*
- Zhong N., Liu J., & Yao Y., (2002). In Search of the Wisdom Web. *IEEE Computer, 35* (11), 27- 31.