Survey Paper on Concept Based Web Search

Komal Rexwal[1], Dr. C.K. Nagpal[2]

[1]Student, Computer Engineering, YMCA University of Science and Technology, Faridabad, India
[2]Professor, Computer Engineering, YMCA University of Science and Technology, Faridabad, India
[1]komalrexwal@gmail.com, nagpalckumar@rediffmail.com

Abstract :

There are many search engines available to deal with keyword based web searches such as Google, Bing etc. But when it comes to concept based search engines there is a lot to be done before we can say that a full-fledged search engine is developed .There are several attempts in developing search engines that deal with some of the aspects of concepts based search. One such attempt is SDE [1] (Search Discover Explore). The main motive of this paper is to study how effective SDE is in dealing with concept based web searches. Also this paper compares SDE with other search engines and see the differences.

Keywords: Concept based web search; exploratory web search; keyword based web search; SDE; system design

1. Introduction

Traditional search engines are keyword based i.e. they treat user query as well as the web documents as a 'bag of words'. A rough working of keyword based search engines is as follows: The documents are indexed according to the words they contain and when a user enters a query, the keywords in the query are matched with the indexed documents and most appropriate documents are returned to user as result. This approach works well if the user has a clear idea of what exactly it wants to retrieve and knows the right keywords that it wants in the retrieved documents. Thus in keyword based search engines the onus is on user to phrase the search query correctly to get appropriate search results.

But if a user does not has a clear understanding of what it wants it will have to use broad terms in its query to describe its requirements. The terms used in such cases represent a concept rather than any particular entity. So, these query terms do not have to be necessarily present in every relevant document rather the presence of entities associated with these query terms should determine the relevance of a document to the user.

The keyword based search engines are not designed to handle such queries and may give unwanted results. To handle such queries we have to perform concept based web search. In concept based web search the focus is on the concepts represented by query terms and the corresponding entities associated with the concepts rather than the keyword. The accuracy of the search result now depends on how accurately the knowledgebase has captured various concepts and related entities and also how accurately the search query is parsed and the probable concepts are extracted out from the query. Thus now the onus of getting the appropriate search results is on search engine whereas previously in keyword based

search engines it was on the user's ability to phrase the query with right keywords. Now the user can give conceptual description of its needs and still will get satisfactory results.

2. Building a concept based Information Retrieval System

2.1 Representation of Resources

For building concept based IR system we have to represent documents as 'bag of concepts' instead of 'bag of word'.
The documents are represented in vector form using Vector Space Model as proposed by salton
[2] dx= (w1x,w2x,….,w|C|x) where wix is weight of concept i in document x and |c| is the total number of concepts. The system described in this paper uses Wikipedia articles as presented by Malo et al, [3].Each document is represented as just a vector of concepts with each concept assigned a weight and concepts that do not occur in a document being assigned a weigh equal to zero.
Milne and Witten [4] algorithm is used to extract concepts from documents in order to convert it into bag of concepts. It has three steps:
I. Candidate selection: In this step the concepts are identified. This is done by generating all possible n-grams and checking the Wikipedia to see which of these n-grams are most frequently used for linking.
II. Concept Resolution: In this step the identified objects in first step are resolved by seeing which Wikipedia article best describe them.
III. Relevance calculation: In this step it is measured that how relevant a concept is to the text. This is done with the help of machine learning algorithms.

2.2 Searching for related resources

The web search should be able to present a user with related resources or documents. There are a number of methods to find relatedness between two documents. The IR system discussed in this paper uses cosine similarity. rel(di,dj)=cos(di,dj)= di*dj ÷ ||di||*||dj||.
The relatedness of two documents varies from 0 to 1 where two documents are completely related if it is 1 and completely unrelated if it is 0.

2.3 Getting recommendation

A good IR system should be able to provide with good recommendations. Recommendations are provided to guide the user toward interesting documents. Here it is assumed that the type of documents that have interested the user in the past will be interesting to the user in the future too.
User's past interest could be gathered by observing user's past activities. User shows his interest in a particular document by doing certain activities like viewing, bookmarking etc. Various activities are assigned weights and the documents are then ranked according to weights.
2.4 Providing with alternative resources
SDE (Search Discover Explore) - the IR system on which are study is based on has this ambitious objective of providing with alternatives so that the user can choose a better option if one is available and the user did not knew about it earlier.
This is done by providing the user with similar documents to the one user is asking for. For example if a user is searching for some tool or software then the system can provide the user with another similar less expensive tool/software in addition to the tool/software the user was looking for.

SDE uses cosine similarity between documents to find similarity between documents. It uses a set of dimensions called contextual dimensions in which the documents can differ. Contextual dimensions could be anything, say, time, cost, location, language etc.

3. Literature Review

According to Guha, McCool and E. Miller [5] actions such as Web Services and the Semantic Web generate a network of distributed machine comprehensible data and present an application called 'Semantic Search' that is built on these things and is an improvement of tradition web search.

Dietze, H.Q. Yu, Giordano, Kaldoudi, Dovrolis and Taibi [6] proposed a general approach to use the treasure of already existing TEL data on the Web. Automated enrichment and interlinking techniques are used to deliver good quality and well-interlinked data for the educational domain.

Reeve and Han[7] in their survey of semantic annotation platforms scrutinizes available Semantic Web annotation platforms that offer annotation and other services, and evaluate their architecture, methods and performance.

Arends, Weingartner, Froschauer, Goldfarb and Merkl [8] adopted concepts from art education and information technology to develop a learning environment for art history and compared artworks along different dimensions without having to rely on textual information.

Giordano, Faro, Maiorana, Pino and Spampinato [9] models a framework for treatment and use of information gathered from different sources of information.

Lops, Gemmis and Semeraro [10] describes the various techniques being employed in current recommender systems and the future scope of the recommender systems.

M. Sahlgren [11] shows that distributional approach to knowledge acquisition are rooted in structural data and can be used to provide knowledge to machines.

Yi and Allan [12] compares utility of different topic models in information retrieval systems.

Günnemann, Derntl, Klamma and Jarke [13] tries to find a method that could process the ever-growing data on web and proposes an interactive text analysis system that exploits dynamic topic modeling to detect the latent topic structure and dynamics in a collection of documents.

4. Comparison

Firstly, we have to choose a topic about which we have to learn about. We are taking ‘cloud computing’ as a topic of interest.

Now, we will try to gather educational resources about cloud computing from Google and SDE and compare the results.

Now we answer the following questions.

TABLE 1. COMPARISON BETWEEN GOOGLE AND SDE

Question	***Google***	***SDE***	***Both Equally***
What platform returns more variety of educational resources?		*SDE*	
Which platform returns more precise results? (Precision: search results are educational		*SDE*	

resources and not other things.)			
Which platform has a better recall? (Recall: you get a big number of educational resources)			*Both*
Which platform better categorizes the results? (Categorization: it means at you know which is the type of a resource in a glance.)		*SDE*	
Which platform provides better disambiguation? (Disambiguation mean how the platforms responds to topics having different meanings)		*SDE*	
Which platform provides better navigation through the space of topics?		*SDE*	
Which platform provides better recommendation?			*Both Equally*

5. Conclusion

In this paper the working of SDE and the comparative effectiveness of SDE against the keyword based search engine (Google) is surveyed. The paper discusses some of the advantages of SDE over Google like the variety of search results, quality of search and representation of search results. Some shortcomings that were felt were also mentioned. It is observed that though SDE gives results from a range of categories some of the results are completely irrelevant. It is observed that –

I. While Google's top search results contain sites like Wikipedia and focuses more on what cloud computing really means, SDE has partitioned its search results under various headings like best results, science journals, sites, documentaries, courses, lectures, events etc. So SDE offers search results in broader categories than Google.

II. The top results of SDE contain high quality documents while Google's search results are comparatively low on quality.

III. SDE gives a small description of the search topic while Google does not.

IV. Every link in SDE search results is accompanied by a picture making SDE visually more attractive.

V. Some of the results in SDE are completely irrelevant to the query.

6. References

1. [1]Roberto Pérez-Rodríguez, Luis Anido-Rifón, Miguel Gómez-Carballa, Marcos Mouriño-García, Architecture of a concept-based information retrieval system for educational resources
2. [2]G. Salton, A. Wong, C.-S. Yang, A vector space model for automatic indexing, Commun. ACM 18(11) (1975) 613–620.
3. [3]P. Malo, A. Sinha, J. Wallenius, P. Korhonen, Concept-based document classification using Wikipedia and value function, J. Am. Soc. Inf. Sci. Technol. 62(12) (2011) 2496–2511.
4. [4] O. Medelyan, I.H. Witten, D. Milne, Topic indexing with Wikipedia, in: Proceedings of the AAAI WikiAI Workshop, vol.1, 2008, pp.19–24
5. [5]R. Guha, R. McCool, E. Miller, Semantic search, in: Proceedings of the 12th International Conference on World Wide Web, ACM, 2003, pp.700–709.
6. [6] S. Dietze, H.Q. Yu, D. Giordano, E. Kaldoudi, N. Dovrolis, D. Taibi, Linked education: interlinking educational resources and the web of data, in: Pro-ceedings of the 27th Annual ACM Symposium on Applied Computing, ACM, 2012, pp.366–371.
7. [7] L. Reeve, H. Han, Survey of semantic annotation platforms, in: Proceedings of the 2005 ACM Symposium on Applied Computing, ACM, 2005, pp.1634–1638.
8. [8] M. Arends, M. Weingartner, J. Froschauer, D. Goldfarb, D. Merkl, Learning about Art History by exploratory search, contextual view and social tags, in: 2012 IEEE 12th International Conference on Advanced Learning Technologies (ICALT), IEEE, 2012, pp.395–399.
9. [9] D. Giordano, A. Faro, F. Maiorana, C. Pino, C. Spampinato, Feeding back learning resources repurposing patterns into the “information loop”: oppor-tunities and challenges, in: 9th International Conference on Information Technology and Applications in Biomedicine, 2009, ITAB 2009, IEEE, 2009, pp.1–6.
10. [10] P. Lops, M. De Gemmis, G. Semeraro, Content-based recommender systems: state of the art and trends, in: Recommender Systems Handbook, Springer, 2011, pp.73–105.
11. [11] M. Sahlgren, The distributional hypothesis, Ital. J. Linguist. 20(1) (2008) 33–54.
12. [12] X. Yi, J. Allan, A comparative study of utilizing topic models for information retrieval, in: Advances in Information Retrieval, Springer, 2009, pp.29–41.
13. [13] N. Günnemann, M. Derntl, R. Klamma, M. Jarke, An interactive system for visual analytics of dynamic topic models, Datenbank Spektrum 13(3) (2013) 213–223.