International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 3, Issue 12, December 2014 ISSN 2319 - 4847

Lucid Gumption Maturation with Query Constellation and Data Retrieval

Dr. Brijesh Khandelwal[1], Dr. S. Q. Abbas[2]

[1]Research Scholar, Shri Venkateshwara University, Merut, UP., India

[2]Research Supervisor, Shri Vinkateshwara University, Merut, U.P. India

Director, Ambalika Institute of Management & Technology, Lucknow, U.P.

ABSTRACT

This research work will be strengthening my ongoing study of Information Retrieval performance Refinement in Deep Web Mining." The information explosion on the Internet has placed high demands on search engines. Yet people are far from being satisfied with the performance of the existing search engines, which often return thousands of documents in response to a user query. Many of the returned documents are irrelevant to the user's need. The precision of current search engines is well under people's expectations. The problem of gumption indistinctness is also one of the reasons of poor performance of the search engines. Deep Web clustering engines have been proposed as one of the solution to the lexical indistinctness issue in Deep Web Information Retrieval. These systems group search results, by providing a cluster for each specific aspect (i.e., meaning) of the input query. Users can then select the cluster(s) and the pages therein that best answer their information needs. The paper tries to discover the success level of query clustering approach for the gumption indistinctness problem in web information retrieval. The extension with annotated framed interface can be exercised for better retrieval of information from deep web too. Furtherance to it, the introduction of logical deep web data retrieval services carries great potential for information retrieval.

Keywords- Query Constellation, Gumption Lucidity, Similarity measure, Web data, Deep Web Access

1.INTRODUCTION

At the retrieval level, traditional approaches are also limited by the fact that they require a document to share some keywords with the query to be retrieved. In reality, it is known that users often use keywords or terms that are different from the documents. There are then two different term spaces, one for the users, and another for the documents. How to create relationships for the related terms between the two spaces is an important issue. The creation of such relationships would allow the system to match queries with relevant documents, even though they contain different terms. This can be made possible with a large amount of user logs. Although keywords are not always good descriptors of contents, most existing search engines still rely solely on the keywords contained in documents and queries to calculate their similarity. This is one of the main factors that affect the precision of the search engines. In many cases, the answers returned by search engines are not relevant to the user's information need, although they do contain the same keywords as the query. The classic approach to information retrieval (IR) would suggest a similarity calculation between queries according to their keywords. However, this approach has some known drawbacks due to the limitations of keywords. In the case of queries, in particular, the keyword-based similarity calculation will be very inaccurate (with respect to semantic similarity) due to the short lengths of the queries. The Traditional Surface Web consists of mostly static content, which is directly inter-linked with static hyperlinks. "Search engines rely on hyperlinks to discover new web pages [6], but static websites are outnumbered by dynamic websites on an extremely large scale and the web has been rapidly deepened [7]. The content as part of dynamic websites are mostly not accessible through static hyperlinks, as these content are dynamically enwrapped into web pages as the response to a structured query submitted through a web query interface. These are intended to be framed by human users to retrieve content from a background database often containing highly relevant content of a specific domain. Common search engines do not reach this part of the web. This is caused by the fact that search engines "typically lack the ability to perform form submissions" [8]. The conceptualization of Logical Framings for information on the web may play a significant role "to absorb information from multiple knowledge sources". This hypothesis can be worked upon, and may be resulting in standards like Resource Description Framework in attributes (RDFa) and Micro data markups like schema.org[9]

2.REVIEW LITERATURE

Although the need for query constellation is relatively new, there have been extensive studies on document constellation, which is similar to query constellation. The first group of related constellation approaches is certainly those that cluster documents using the keywords they contain. In these approaches, in general, a document is represented as a vector in a vector space formed by all the keywords [1]. Researches have been concerned mainly with the following two aspects:

- similarity function
- algorithms for the constellation process

Extending the concept for retrieval and indexing of Deep Web content have been addressed from different perspectives in the past. The effort has mostly focused specific applications to discover, retrieve and index structured data from the Deep Web. We can always describe the different types of structured data in the context of the varying search tasks that we can strive to support over them. This may include special emphasis on the automatic web query interface interpretation. Common approaches focusing on exposing Deep Web content can also be classified to surfacing and virtual integration approaches at the same time. If two documents are judged relevant to the same query, then there are reasons to believe that these documents talk about the same topic, and therefore can be included in the same cluster. Incorporating user judgments, in this way, may solve some of the problems in using keywords. However, in a traditional IR environment, the amount of relevance feedback information is too limited to allow for a reasonable coverage of documents. In the Web environment, the choice of a particular document from a list by a user is another kind of cross-reference between queries and documents. Although it is not as accurate as explicit relevance judgment in traditional IR, the user’s choice does suggest a certain degree of “relevance” of that document to his information need. In fact, users usually do not make the choice randomly. Similar ideas have been used in some work in IR. [2] Try to evaluate the quality of an IR system for a cluster of queries on different document collections. Their goal is to determine an appropriate method for database merging according to the quality estimation. The assumption used in query constellation is that if two queries retrieve many documents in common, they are on the same topic. The most similar work is that of [3], which exploits the same hypothesis for document and query constellation. However, while exploiting cross-references, [3] reject the use of the contents of queries and documents. They consider that keyword similarity is totally unreliable. We believe that content words provide some useful information for query constellation that is complementary to cross-references. Therefore, our approach tries to combine both content words and cross-references in query constellation. In contrast to it, using query keywords directly is not reliable due to their short length and word indistinctness. In [4], the top documents retrieved by a search engine for each query are considered as the data source of the query and a hierarchical constellation algorithm is applied. Based on the click-through data, in [5], a bipartite graph of queries and documents is constructed and then a graph based agglomerative iterative constellation method is applied to merge vertices of graph continually until a termination condition reaches. Similar approach is also followed by [10] and found good results. Going beyond the clicked URLs in the search engine query and browsing logs [11], developed one algorithm to use query keywords and the clicked documents to estimate the similarity between queries. [12] Developed a hybrid similarity measure feature for the constellation approach. In [13] [14], a query is assumed ambiguous based on the underlying document collection. [13] Introduced a query lucidity method by considering part-of-speech patterns of the query’s context terms. For each query, [13] analyzed its context in the data collection and classify each occurrence of the query to a question which reflects its context type. Indistinctness is resolved by choosing a question from the generated question list. Such a method suffers from scalability issues and is not applicable to the Web environment. [15]

3. QUERY CONSTELLATION APPROACHES IN WEB

A. Using Content words and End User Feedback

In particular, we make use of the cross-references between the users' queries and the documents that the end users have chosen to read. The approach follows the hypothesis that there is a strong relationship between the queries and the selected documents (or clicked documents). It is based on the following principle - If two queries lead to the same document clicks, then they are similar or related. Document clicks are comparable to user relevance feedback in a traditional IR environment, except that they denote implicit and not always valid relevance judgments. This principle is used in combination with the traditional approaches based on query contents. [9] Content keywords are the words except function words included in a stop-list. All the keywords are stemmed using the Porter's algorithm [16]. Following formula is used to measure the content similarity between two queries,

$$Sim_{keyword}(p,q) = \frac{KN(p,q)}{Max(kn(p),kn(q))}$$

where *kn(.)* is the number of keywords in a query, *KN(p, q)* is the number of common keywords in two queries.

A first feedback-based similarity considers each document in isolation. This similarity is proportional to the number of the clicked individual documents in common for two queries p and q as follows:

$$Sim_{click}(p,q) = \frac{RD(p,q)}{Max(rd(p),rd(q))}$$

where *rd(.)* is the number of clicked documents for a query and *RD (p, q)* is the number of document clicks in common. A very similar study [3] has been carried out recently. However, that study rejects the use of content words and relies solely on document clicks to cluster queries. We think that both query contents words and the corresponding document clicks can partially capture the users' interests. Therefore, it is better to use both. A simple way to do it is to combine both measures linearly as follows:

$$Sim_{comp} = \alpha * Sim_{content} + \beta * Sim_{feedback}$$

There is an issue concerning the setting of parameters α and β. In our current implementation, these parameters are to be set manually by the users in order to obtain different behaviors.

B. Hybrid approach

The content-based query constellation approach groups different queries with the same or similar keywords. However, a single query term can represent different information needs. The results-based approach determines the relationship between queries using the results returned by a search engine. This method uses more contextual information for a given query. However, the same document in the search results listings might contain several topics, and thus queries with different semantic meanings might lead to the same search results. While the content-based approach may not be suitable for query constellation by itself, query terms have been shown to have the ability to provide useful information for constellation [11]. Therefore, we believe that the content-based approach can augment the results-based approaches and compensate for the indistinctness inherent in the latter. Hence, unlike [17], we assume that a combination of both methods will provide more effective constellation results than using each of them individually. Based on this hypothesis, we define a hybrid similarity measure as [18]:

$$Sim\_hybrid(Q_i,Q_j) = \alpha * Sim\_result(Q_i,Q_j) + \beta * Sim\_\cos ine(Q_i,Q_j)$$

where α and β are parameters assigned to each similarity measure, with $\alpha+\beta=1$. Here, α and β represent varying levels of contribution a particular approach (results-based or content-based respectively) has in determining the similarity between queries.

C. Incremental approach

The system incorporates implicit indistinctness resolution method based on query-oriented document clusters. In the system, a query in Korean is first translated into English by looking up dictionaries, and documents are retrieved based on the vector space retrieval for the translated query. For the top-ranked retrieved documents, document clusters are incrementally created and the weight of each retrieved document is re-calculated by using clusters with preference. This phase is the core of our implicit indistinctness resolution method. While synonyms can improve retrieval effectiveness, terms with different meanings produced from the same original term can degrade retrieval performance tremendously. At this stage, we can apply statistical indistinctness resolution method based on mutual information. For the query, documents are retrieved based on the vector space retrieval method. This method simply checks the existence of query terms, and calculates similarities between the query and documents. The query-document similarity of each document is calculated by vector inner product of the query and document vectors:

$$simD(q,d) = \sum_{i=1}^{t} w_{qi}.w_{di}$$

where query and document weight, *qi w* and *di w* , are calculated by *ntc-ltn* weighting scheme which yields the best retrieval result in [17] among several weighting schemes used in SMART system [18]. As the translated query can contain noises, non-relevant documents may have higher ranks than relevant documents. It should be noted here that the static global constellation is not practical in the current setup, because it takes much computational time and the document space is too sparse [19] for the comparison of static and dynamic constellation). The clusters made are based on incremental centroid method. There are a few variations in the agglomerative constellation method. The agglomerative centroid method joins the pair of clusters with the most similar centroid at each stage [20]. Incremental centroid constellation method is straightforward. The input document of incremental constellation proceeds according to the ranks of the top-ranked N documents resulted from vector space retrieval for a query. Document and cluster centroid are represented in vectors. For the first input document (rank 1), create one cluster whose member is itself. For each consecutive document (rank 2, ..., N), compute cosine similarity between the document and each cluster centroid in the already created clusters. If the similarity between the document and a cluster is above a threshold, then add the document to the cluster as a member and update cluster centroid. Otherwise, create a new cluster with this document.

Similarities between the clusters and the query, or query-cluster similarities, are calculated by the combination of the query inclusion ratio and vector inner product between the query vector and the centroid vectors of the clusters.

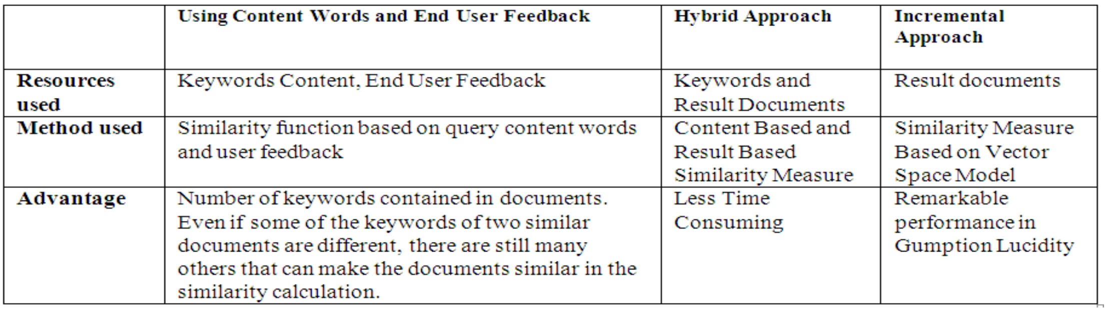
$$simC(q,c) = \frac{c_q}{q}.\sum_{i=1}^{t} w_{qi}.w_{ci}$$

Where $|q|$ is the number of terms in the query, $|cq|$ is the number of query terms included in a cluster centroid, $|cq|/|q|$ is the query inclusion ratio for the cluster. The documents included in the same cluster have the same query-cluster similarity. Cluster preferences are influenced by the query inclusion ratio, which prefers the cluster whose centroid includes more various query terms. Thus incorporating this information into the weighting of each document means adding information which is related to the behavior of terms in documents as well as the association of terms and documents into the evaluation of the relevance of each document; it therefore has the effect of indistinctness resolution.

Overall comparison of all approaches in terms of resources and methods used along with the advantages

Keyword-based document constellation has provided interesting results. One contributing factor is the large number of keywords contained in documents. Even if some of the keywords of two similar documents are different, there are still many others that can make the documents similar in the similarity calculation. However, since queries, especially the queries submitted to the search engines, typically are very short, in many cases it is hard to deduce the semantics from the queries themselves. Therefore, keywords alone do not provide a reliable basis for constellation queries effectively. [9] In the hyperlink approaches, document space and query space are still separated. The question is whether it is possible to exploit the cross reference between documents and queries in query/document constellation. By cross-reference, we mean any relationship created between a query and a document. The intuition of using cross-references is that similarity between documents can be transferred to queries through these references, and vice versa. [9] See Table 1.

TABLE 1: OVERALL COMPARISON OF ALL APPROACHES



	Using Content Words and End User Feedback	**Hybrid Approach**	**Incremental Approach**
Resources used	Keywords Content, End User Feedback	Keywords and Result Documents	Result documents
Method used	Similarity function based on query content words and user feedback	Content Based and Result Based Similarity Measure	Similarity Measure Based on Vector Space Model
Advantage	Number of keywords contained in documents. Even if some of the keywords of two similar documents are different, there are still many others that can make the documents similar in the similarity calculation.	Less Time Consuming	Remarkable performance in Gumption Lucidity

5.DATA RETRIEVAL APPROACHES IN DEEP WEB

For the targeted data retrieval through possible interface of query constellation especially of dynamic Deep Web content, the need of an efficient and automatic approach is mandatory. Therefore, the attention needs to be set to these challenges: content providing service Detection, Incantation & Implementation and the Composition. By meeting these challenges we will ensure the identification of appropriate web query interfaces providing access to relevant content (Detection), the appropriate query mapping and query submission (Incantation & Implementation) and the service interoperability (Composition) as described in [8]. Common approaches for Deep Web Data Retrieval focus these challenges from the data retrieving services perspective. The conceptual idea being introduced in this section focuses these challenges from the information providing services viewpoint.b.0 Another dimension of such information retrieval system with respect to structured interface contains clusters of connected attributes that may affect each other. The first select attribute as part of the designated cluster defines the relation to the other clusters. The second select attribute as part of this cluster defines the input attribute domain and may restrict the input attribute range of the input attribute that is part of the referred clusters (Figure-1). More complex examples may establish that Logical meaning behind a Logical Deep Web Data Retrieval System (LDWDRS) interface might be quite complex and automated form understanding approaches may quickly reach their limits. Especially the automated detection of connected attributes and the detection of complex relations within clusters might be the most difficult but uncommon successful part.

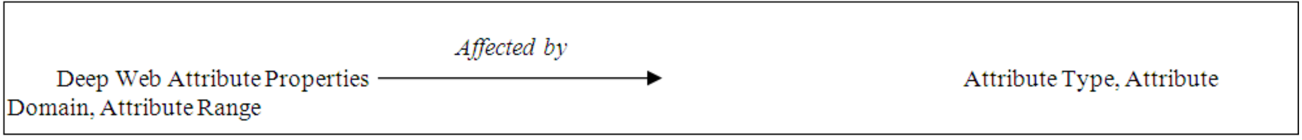


Figure 1 Clusters of connected attributes properties in Data Retrieval System

5.CONCLUSION

In this paper we had tried to find out the role of query constellation in gumption lucidity with respect to web along with a study of having scope of conceptualizing a structured framework through query constellation for retrieval of data from deep web as an extension to it. As discussed in section IV we can justify the role of gumption constellation in WSD. Gumption Indistinctness is one of the core problems of web information retrieval. Gumption Indistinctness deteriorates the performance of the web information retrieval. The paper concludes that constellation approach is helpful in lucidity of word gumptions. The researchers showed high level of performance level in using constellation approach for WSD. The paper also outlined the various approaches for constellation and also their benefits and drawbacks. At the same time, in context to deep web data retrieval, a variety of current data retrieval mechanisms and prescribed structured form understanding systems endeavor to study logical deep web data retrieval systems' structured interfaces automatically by concentrating the Deep Web Data Retrieval challenge from the retrieving services angle. The studied system follows the open knowledge sharing model as part of the Logical Web vision from Berners-Lee et al. [21]. This is based on the assumption, that the information provided on websites is planned to be retrieved by various services.

REFERENCES

[1] G. Salton And M. J. Mcgill, Introduction to Modern Information Retrieval. McGraw-Hill New York, NY, 1983.

[2] E. Voorhees, N. K. Gupta, and, B. Johnson-Laird, Learning collection fusion strategies. In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, New York, NY, 172–179, 1995.

[3] D. Beeferman and A. Berger, Agglomerative clustering of a search engine query log, In Proceedings of the 6th CMSIGKDD International Conference on Knowledge Discovery and Data Mining, (August). Acm Press, New York, NY, 407–416, 2000.

[4] S.-L. Chuang and L.-F. Chien, Towards automatic generation of query taxonomy: a hierarchical term clustering approach, In Proceedings of 2002 IEEE International Conference on Data Mining, (ICDM), 2002

[5] D. Beeferman and A. Berger, Agglomerative clustering of a search engine query log, In Proceedings of the sixth ACM SIGKDD, pages 407--415, 2000.

[6] Madhavan, J., Ko, D., Kot, L., Ganapathy, V., Rasmussen, A., and Halevy, A. Google's deep web crawl. Proceedings of the VLDB Endowment 1, 2 (2008), 1241{1252.

[7] He, B., Patel, M., Zhang, Z., and Chang, K. C.-C. Accessing the deep web. Communications of the ACM 50, 5 (2007), 94{101.

[8] Arne Martin klemenz, Klaus Tochtermann, Semantification of Query Interface to Improve Access to Deep Web Content, SDA 2013, 3rd International Workshop on Semantic Digital Archives.

[9] B. Khandelwal and Parul Verma, Query Clustering and its Application in Word Sense Disambiguation (WSD), International Journal of Emerging Trends & Technology in Computer Science (ISSN 2278-6856), Volume 1, Issue 2, July-August 2012.

[10] J. Yi and F. Maghoul, Query Clustering using Click-Through Graph, In Proceedings of the 18th international conference on World wide web, New York, USA, 1056-1057,2009.

[11] J.-R Wen., J.-Y. Nie, and H.-J Zhang. Query Clustering Using User Logs, In ACM TIOS, vol. 20, no. 1, pp. 59-81, 2002.

[12] L. Fu., D.H. Goh and S. Foo, Query clustering using a hybrid query similarity measure, WSEAS Transaction on Computers, 3(3), 700-705, 2004

[13] J. Allan and H. Raghavan, Using part of speech patterns to reduce query indistinctness, In Proceedings of the 25th annual international ACM SIGIR, pages 307--314, 2002.

[14] S. Cronen-Townsend and W. B. Croft Quantifying query indistinctness. In Proceedings of HLT, pages 94--98, 2002.

[15] F. Liu, C. Yu, and W. Meng, Personalized Web search by mapping user queries to categories, In Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM '02). USA, 558—565, 2002

[16] M. Porter, An algorithm for suffix stripping, Program, Vol.14, No. 3, pp. 130-137, 1980.

[17] K.S. Lee, Y.C. Park, and K.S. Choi, Re-ranking model based on document clusters. Information Processing and Management, Vol. 37, No. 1, pp. 1-14., 2001

[18] G. Salton, Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, Reading, Pennsylvania, 1989

[19] P.G. Anick and S. Vaithyanathan, Exploiting Clustering and Phrases for Context-Based Information Retrieval. In Proc. of 20th ACM SIGIR Conference (SIGIR'97), 1997

[20] W.B. Frakes, and R. Baeza-Yates, Information Retrieval: data structures & algorithms. New Jersey: Prentice Hall, pp.435-436, 1992

[21] Berners-Lee, T., Hendler, J., Lassila, O., et al. The Logical web. Scienti_c American 284, 5 (2001)

AUTHOR

Dr. Brijesh Khandelwal did MCA from Lucknow University in year 1994. In 2001, he became Sun Certified Programmer. In 2007 he did PhD (Appllied Economics) from Lucknow University. He did MBA in 2010 from Punjab Technical University. In 2010 he also became licentiate in Life Insurance from Insurance Institute of India, Mumbai.

Dr. S. Q. Abbas has more than 23 years of experience in Academic and Administration. Currently, he is DG at Ambalika Institute of Management& technology, Lucknow. Dr. Abbas has rich & diverse experience in academia and is Visiting Professor at various universities & colleges. He has several publications in International/ National Journals & Conferences. He supervised many candidates of Ph.D & M.Phil. He is in advisory board and reviewer of various Int. & National Journals.