Global Journals LaTeX JournalKaleidoscope™

Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

A New Modified Collection Selection Algorithm using Optimal Term Weight for Web based Applications

K. S. Niraja[1], B. Ramana Reddy[2] and B. Ramana Reddy[3]

[1] Muffakhamjah College of Engineering

Received: 10 December 2015 Accepted: 4 January 2016 Published: 15 January 2016

Abstract

As the number of electronic data collections available on the internet increases, so does the difficulty of finding the right collection for a given query. Often the first time user will be overwhelmed by the array of options available, and will waste time hunting through pages of collection names, followed by time reading results pages after doing an adhoc search. Collection selection using optimal weight methods try to solve this problem by suggesting the best subset of collections to search based on a query. This is of importance to fields containing large number of electronic collections which undergo frequent change, and collections that cannot be fully indexed using traditional methods such as spiders. This paper presents a solution to these problems of selecting the best collections and reducing the number of collections needing to be searched.

***Index terms*—** singular value matrix(s), term matrix (u), collection matrix (v).

Introduction he 21st century is the age of Internet and World Wide Web. The Web revolutionizes the way we gather, process, and use information. At the same time, it also redefines the meanings and processes of business, commerce, marketing, finance, publishing, education, research, development, as well as other aspects of our daily life [1].

Modified Collection selection is the selection of an optimal weight subset of collections from a large set of collections for the purpose of reducing costs associated with Distributed Information Retrieval. The goal of modified collection selection is to make searching multiple collections appear as seamless as searching a single collection. Another requirement of a modified collection selection using optimal term weighting system is to learn which collections contain relevant information and which collections contain no relevant information. This reduces the number of overall search requests needed. If only a small high quality subset of the available collections is searched then savings can be made in time, bandwidth, and computation [4]. Web based collection selection is significant because as the internet grows the number of internet based collections grows. It is now impossible to anually track and index all collections as they number in the thousands. This method will enable users to choose the best collections for their needs without having to sift through irrelevant collections. Collection selection optimal term method reduces expenses, increasing search speed, learning to adapt to change in the search environment, using ontology to increase precision, and learning to adapt to the users preferences.

The paper is organized as follows. Section 2 discusses main difference between traditional method in web based collection and optimal term weight method for collection selection Section 3 presents application of the approach. Conclusion presents main features of the system that help fulfill fundamental demands of the intelligent Web's design and development II.

1 Modified Collection Selection

Modified Collection Selection using optimal term is the selection of an optimal set of information sources from a large set of information sources. An information source can be a Web interface, a standard relational collection, a file, a search engine, or any other textual representation of information. Collection Selection aims to be efficient with respect to bandwidth and computation, and decreases both resource usage and time taken to return a set of results for a query. Well planned collection selection can have a large influence on the efficiency of a query. Collection selection is significantly different to document selection in a number of areas. Collection selection uses different methods to document selection for scoring items relevance [3]. Document selection commonly uses a binary relevance value, which collection selection cannot use. Instead collection selection must use a floating point number to represent relevance.

Collection selection also differs from document selection in that it uses different ways of calculating term weighting. (terms distributed across all documents in a collection are worth more than terms clustered in one document of a collection) Another difference between collection selection and document selection is that different content selection methods are needed, with Web based collection selection commonly using partial collection sampling, and document selection using full document indexing. These differences mean that collection selection using optimal term requires a significantly different approach to document selection III.

2 Modified Collection Selection Algorithm

In this section, we give the details of our collection selection algorithm. The inputs of the algorithms include a query, a selected set of terms (key words), and a set of sample documents from each collection. a) Algorithm 1. Calculate the term-collection matrix A where we view the query as a new collection.

2. Use singular value decomposition. U ?V T = A 3. Sort the collections according to the values in the query row in the matrix V T 4. Use the threshold to calculate a rank of collections. 5. After ranking the collection we need to find the optimal term weight to find the relevant pages which are more appropriate. Term-collection matrix is created, adding the query to the matrix in the form of a new (small) document column. Negative weights can be given to terms that are not to be returned in the query. Applying Singular Value Decomposition to the matrix returns a term matrix (U), a Singular Value matrix(S), and a collection matrix (V). For every search performed, the user will give the top n collections (n is currently 10) a floating point precision ranking in the range of 0 to 1.

The higher the ranking the more precise the results. After training run of (say) twenty searches collection matrix and the latent statistical relationships between collections computed [5].The returned values are a score for each collection, with zero being not relevant and one being most relevant. This will find relationships existing between collections that are not immediately obvious, and will result in a more personalized search which will over time learn the user's preferences.

IV.

3 Conclusions

A solution to the Web Based Collection Selection problem has been presented, and preliminary results indicate that the technique is suited to the task of selecting the most relevant collections and learning user preferences in collections. The approach uses short queries and is thus suitable for use on the Web. This approach also reduces the need for ontologies and thesaurus. With some modification, this collection selection method is suitable for traditional information retrieval systems across servers and databases. A problem is that these systems do not rank the data before returning it. This could be solved using simple sampling techniques that would grab a representative sample of the collection, rank it, then compare it across collections. As the number of collections indexed grows, so does the number of terms and the size of the matrix.

However in this research, only the top n most representative documents from each collection are sampled so it is possible to compare hundreds of collections in a reasonable time if n is small. Due to the time expense of writing screen scraping applications for web based collections and comparing the results to human rankings of the documents in the collections, the researchers were unable to perform large scale tests of the methods presented in this research. Work still needs to be done to on the optimal sample size taken from each collection.

[References Références Referencias] , References Références Referencias .

[Berry et al. ()] *Matrices, vector spaces, and information retrieval*, M W Berry , Z Drmac , E R Jessup . 1999. Society for Industrial and Applied Mathematics. 41 p. .

[Craswell et al. ()] 'Server selection on the World Wide Web'. N Craswell , P Bailey , D Hawking . *Proceedings of the fifth ACM conference on Digital libraries*, (the fifth ACM conference on Digital librariesSan Antonio, Texas, United States) 2000. ACM Press. p. .

[Callan et al. ()] *The effects of query-based sampling on automatic database selection algorithms*, J Callan , A L Powell , J C French , M Connell . CMU-LTI-00-162. 2000. Language Technologies Institute, School of Computer Science, Carnegie Mellon University (Technical Report)

[US) Guidelines Handbook Global Journals Inc ()] 'US) Guidelines Handbook'. www.GlobalJournals.org *Global Journals Inc* 2016.

[King and Li] *Web Based Collection Selection Using Singular Value Decomposition School of Software Engineering and Data*, John King , Yuefeng Li . Australia. Communications Queensland University of Technology QLD 4001

[Zhong et al. (2002)] N Zhong , J Liu , Y Yao . *Search of the Wisdom Web*, November 2002. 35 p. .