Relevance Feedback versus Web Search Document Clustering

Mansaf Alam
Jamia Millia Islamia
144 PUBLICATIONS **1,543** CITATIONS
SEE PROFILE

Kishwar Sadaf
Majmaah University
12 PUBLICATIONS **151** CITATIONS
SEE PROFILE

Relevance Feedback versus Web Search Document Clustering

MansafAlam
Dept. of Computer Science
JMI, New Delhi,INDIA

KishwarSadaf
Dept. of Computer Science
JMI, New Delhi,INDIA

Abstract- The performance of an IR system is deteriorated by factors including short and vague queries put up by the users, ever increasing volume of documents on the web, users not knowing their exact information need etc. Relevance feedback (RF) and web search document clustering are techniques to improve the performance of an Information Retrieval (IR) system.Relevance feedback provides a method to get more relevant search result from an IR system using documents that are marked relevant by the user as a feedback to reformulate query. This refined query is then used to retrieve the documents. In document clustering approach, the search result is divided into thematic groups where documents of one group are similar to each other and dissimilar to the documents of other groups. This paper presents a report on the effectiveness of relevance feedback technique as compared to document clustering in context of web information retrieval and why document clustering is the most preferred approach.

Keywords- Relevance Feedback, Document Clustering, IR, web search.

I. INTRODUCTION

Finding relevant search result from web is one the most researchable area in the field of computer science. An experienced user who knows about his/her actual information need, may be able to quickly locate the desired and relevant result but a novice user with ill-structured query may roam around the millions of documents with no success. This happens because a general search engine or web IR system suffers from deficiencies. For example, a simple and small query may return millions of documents and of which only few are marginally relevant and to get to those document is not also effortless if the same is not given on the first display of the system. The other problem is polysemy, where a document is highly relevant but doesn't show or buried deep in the search result due to lack of words that represent query. RF and document clustering are techniques to improve retrieval process. Relevance feedback is an information retrieval process where a query is reformulated using previously retrieved documents [1]. It focuses on how to refine ill-structured or ambiguous queries. In a typical relevance feedback system, a user explicitly marks the initial retrieved documents as relevant or irrelevant. The query is then reformulated using these documents and more relevant documents are retrieved. But that is not always the case as we will see later in this paper. Document clustering is a technique to groupdocuments by topic into different clusters. The clustering of web search documents, assist the users to quickly locate what they are looking for. The users don't have to reformulate the query, but have to simply click on the topic which most closely or accurately defines their information need. Clustering of web information retrieval result or documents is an active research area. Since relevance feedback is a after retrieval process, we consider only those document clustering techniques that are applied on the retrieved result. In this paper we try to find out what makes RF lag behind document clustering in web information retrieval scenario.

II. RELEVANCE FEEDBACK (RF)

The ill-formed query degrades the performance of the retrieval. Reformulating query is the process of adding additional terms to the query according to user's information need. Relevance feedback is basically a query expansion technique. Feedback given to an IR system can be categorized as:

- Relevance Feedback(RF)
- Pseudo Relevance Feedback (PRF)
- Indirect Relevance Feedback

The typical relevance feedback (RF) procedure consists of following steps:

- The user fires a query.
- The IR system responds with a set of documents.
- The user has to mark some retrieved documents as relevant or non-relevant.
- The query is refined using the user's judgment.
- Using this query, a new improved result is presented to the user.

The above process may take place iteratively. The performance of the system basically depends on the underlying retrieval model being employed. Modern day IR systems provide feedback in the form of "similar" or "related" options. These types of feedback are explicit. A user explicitly indicates his/her judgement. Feedback can be implicit where there is no extended interaction between user and the system. The system keeps track of the documents a user selects or skips. If a user jumps past few documents and hop to a document that is not in top, means that top documents are not relevant and hence are not used in query reformulation. Pseudo relevance feedback also known as Blind Relevance Feedback is a type of query expansion technique where auser is not involved in the judgement process directly or indirectly. Introduced by Croft et al. [2], in this method, no knowledge about the relevance/irrelevance of a document is provided. The top k documents that are initially retrieved are assumed as relevant. The query is the expanded using these documents' words. In all of the above technique, a feedback set of documents are formed. The words or terms of the documents are analysed to select words that describes the feedback set most accurately.

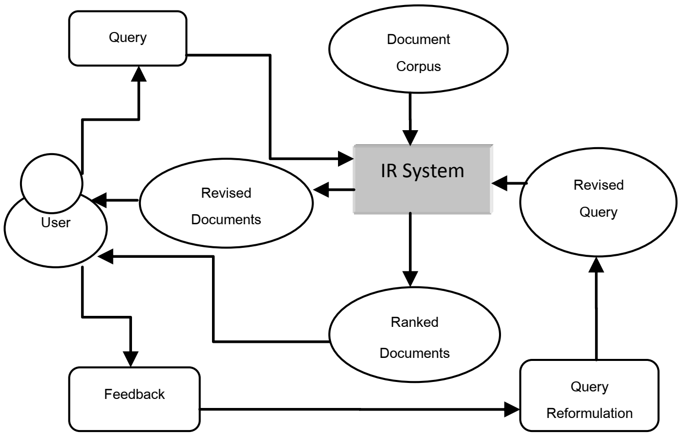


Fig.1. Relevance Feedback Architecture

Rocchio's [3] algorithm is the first formalization of the RF model. It models a way of incorporating relevance feedback information into the vector space model [4]. Query and documents are first turned into vectors. The underlying theory of this approach is to move query vector towards the centroid of the relevant documents vectors and away from the centroid of irrelevant documents. Given the query $\vec{q}$ and information about relevant and non-relevant document sets D_r and D_{nr}, the algorithm maximizes similarity between query andrelevant documents while minimizing similarity between query and non-relevant documents as:

$$\overrightarrow{q}=\arg\max[sim(\overrightarrow{q}, D_r)-sim(\overrightarrow{q}, D_{nr}\ )] \qquad (1)$$

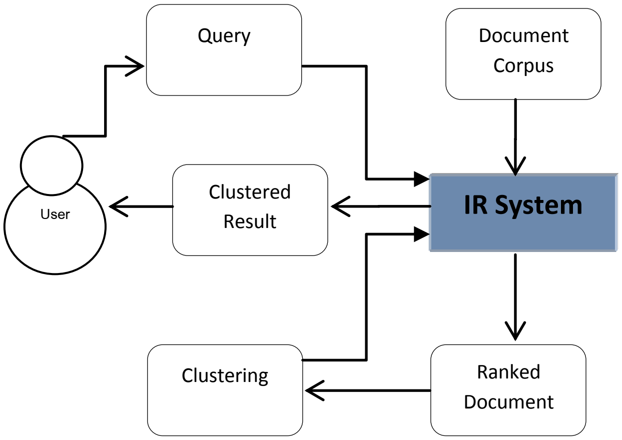
The new query vector is the concatenation of original query and the terms that represent the feedback documents justly. [5] Provides a comprehensive study on the usage of relevance feedback approach in information access systems.In [6],authors provide a model that considers positive as well as negative feedback. Authors in [7] use collaborative filtering for pseudo relevance feedback.

III. DOCUMENT CLUSTERING

Document Clustering is method of bringing documents together to form groups or clusters where documents of one cluster are similar to each other whereas dissimilar to the documents of other clusters. It is being sought as a powerful technique to improve the information retrieval process. This process requires no involvement of the user. The documents that are returned in a response to a query are grouped into different clusters. However, small or ambiguous is a query, clustering can produce different clusters for possible meanings of the query words. If a query contains "apple" keyword, a clustering based IR system returns documents clusters related to possible meanings of "apple": the company Apple Inc. or the fruit apple. A user can easily locate the information he or she is looking for without having to give feedback or specifying the query as "apple systems" or "apple fruit".

Document clustering is an unsupervised machine learning technique. This makes it feasible for web information retrieval. Document clustering in the context of web IR can be categorised as pre-retrieval clustering and post-retrieval clustering. In pre-retrieval clustering, documents are first identified using query terms and then clustered. It involves large corpus of documents. In post-retrieval approach, using query terms, document snippets that are returned by the web IR system are clustered. The corpus here is relatively low as compared to pre-retrieval technique. Whether pre or post, clustering can be of two types: agglomerative and flat [8]. Agglomerative clustering methods group the documents into a hierarchical tree structure. The procedure starts by placing each document into a distinct cluster. Then two most similar documents are merged. This process iterates. A hierarchical tree structure of the clusters is presented to the user. The hierarchy of clusters can also be created in top-down order, which is known as the divisive approach. It starts with all the documents assigned to one cluster and iteratively divides a cluster into smaller clusters until a certain termination threshold is achieved. On the other hand, document partitioning methods decompose a document corpus into a given number of disjoint clusters. The number of partitions must be decided in advance. Partitioning methods can also generate a hierarchical structure of the document corpus by iteratively partitioning a large cluster into smaller clusters. K-means clustering is one the most famous and common flat clustering techniques.

A common technique used by conventional clustering search engine is to cluster document snippets rather than using complete documents. A document snippet is a short description of a document, contains query words and indicates the user its contents. It has been a matter of research that what drive yield the best result: document snippets or whole documents.An important document snippet-based clustering, Suffix Tree Clustering (STC), is based on the Suffix Tree Document (STD) model which was proposed by [9]. Many other works based on STC are available in literature. [10] presented a study where different web search document clustering techniques are reviewed. In [11], authors proposed a method. First, labels for clusters are defined using the document snippets and then documents are assigned to these clusters according to their similarity with the labels. In [12], Mecca et al utilizes the powerful mathematical function Singular Value Decomposition (SVD) on documents returned by the search engine as a complete instead of document snippets. Their algorithm has been integrated with Noodles search engine. Search result clustering whether using snippets or whole document gives words result in a concise and meaningful way.

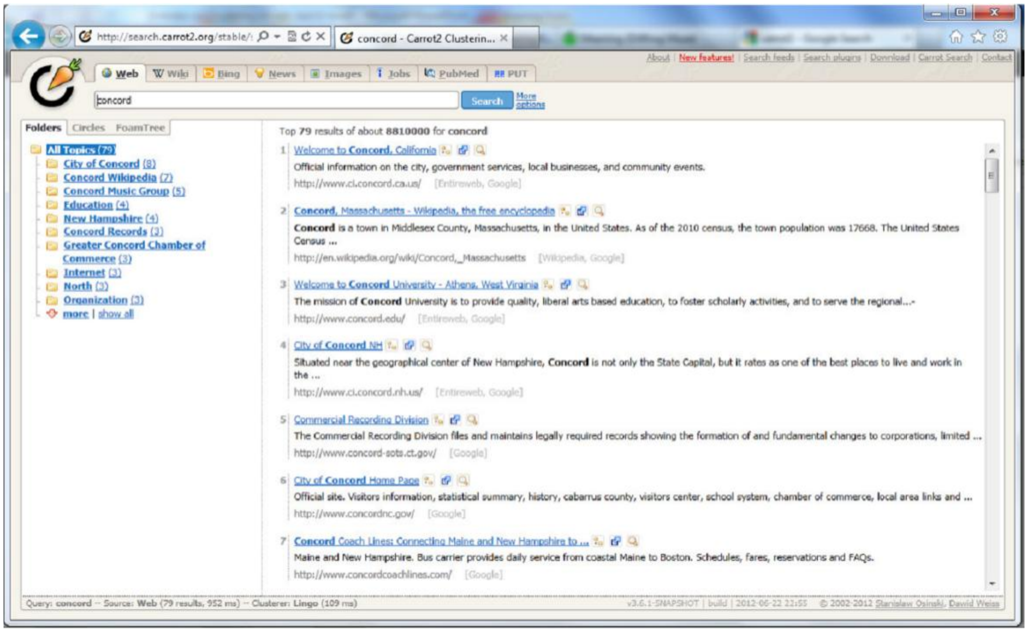


IV. DOCUMENT CLUSTERING OR RELEVANCE FEEDBACK

Document clustering and RF are the techniques to enhance the performance of retrieval process. We examine the reasons why

RF is not widely pertinent as document clustering in context of web IR. An ideal Information Retrieval system performs effectively and efficiently. An effective and efficient IR systemgives relevant result without delay. Since relevant feedback is after-retrieval process, we consider clustering of documents post retrieval.Relevance feedback has not been used full-fledged in web search engines. Although it is being used in image search, but has become obsolete in textual search. The reason of this The factors that make relevance feedback almost out of use are:

- The user must have sufficient knowledge to be able to put a query which describes his information need. Most searchers are novice. We cannot hope to get an ideal query by a novice user.
- Misspelling, cross language IR, mismatch between user's vocabulary and document collection vocabulary (Synonymy).
- Long query resulted from feedback severely degrades the IR performance.
- User does not want to prolong the search process.
- Several iterations decrease the relevance feedback improvements.
- Pseudo RF has one major drawback-Query drift. If the initial top k document set contains few or no relevant document, the RF will add words to query


http://search.carrot2.org/stable/
concord - Carrot2 Clusterin...
About | New features! | Search feeds | Search plugins | Download | Carrot Search | Contact
Web
Wiki
Bing
News
Images
Jobs
PubMed
PUT
concord
Search
More options
Folders
Circles
FoamTree

All Topics (79)
City of Concord (8)
Concord Wikipedia (7)
Concord Music Group (5)
Education (4)
New Hampshire (4)
Concord Records (3)
Greater Concord Chamber of Commerce (3)
Internet (3)
North (3)
Organization (3)
more | show all

Top 79 results of about 8810000 for concord

1 Welcome to Concord, California
Official information on the city, government services, local businesses, and community events.
http://www.ci.concord.ca.us/ [Entireweb, Google]

2 Concord, Massachusetts - Wikipedia, the free encyclopedia
Concord is a town in Middlesex County, Massachusetts, in the United States. As of the 2010 census, the town population was 17668. The United States Census ...
http://en.wikipedia.org/wiki/Concord,_Massachusetts [Wikipedia, Google]

3 Welcome to Concord University - Athens, West Virginia
The mission of Concord University is to provide quality, liberal arts based education, to foster scholarly activities, and to serve the regional...-
http://www.concord.edu/ [Entireweb, Google]

4 City of Concord NH
Situated near the geographical center of New Hampshire, Concord is not only the State Capital, but it rates as one of the best places to live and work in the ...
http://www.ci.concord.nh.us/ [Entireweb, Google]

5 Commercial Recording Division
The Commercial Recording Division files and maintains legally required records showing the formation of and fundamental changes to corporations, limited ...
http://www.concord-sots.ct.gov/ [Google]

6 City of Concord Home Page
Official site. Visitors information, statistical summary, history, cabarrus county, visitors center, school system, chamber of commerce, local area links and ...
http://www.concordnc.gov/ [Google]

7 Concord Coach Lines: Connecting Maine and New Hampshire to ...
Maine and New Hampshire. Bus carrier provides daily service from coastal Maine to Boston. Schedules, fares, reservations and FAQs.
http://www.concordcoachlines.com/ [Google]

Query: concord -- Source: Web (79 results, 952 ms) -- Clusterer: Lingo (109 ms)
v3.6.1-SNAPSHOT | build | 2012-06-22 22:55 © 2002-2012 Stanislaw Osinski, Dawid Weiss

Fig. 3 Carrot2 Clustering Engine

which are not relevant. Hence the whole result is flawed.

In [13], authors experimented that PRF performs well only in clinical environment. [14] proposed a method where to avoid query drift, only subset important terms from the feedback set are used. All this work only when the initial query put up by the user defines his/her actual information need. Relevance feedback is basically a recall enhancing method [15]. Web search users are rarely troubled with getting higher recall. Spink et al [16] conducted a study on Excite web search engine. They showed that only around 4% of users opted for relevance feedback. Only about 70% of users browsed the first page of the result and left the page. Web search document clustering is an unsupervised machine learning technique.The system has no prior knowledge of what clusters will be produced. It solely depends on the query. The document clustering approach addresses the short, ambiguous query problem. The user only has to put the query. No further user interaction is needed.The user get to see the whole result set in the form of clusters on the first display itself.User can select a cluster of documents that rightly defines his or her information need.

Clusters should have meaningful labels. If a label does not describe its contents smartly, no matter how relevant the documents, a user would not select it. Kartoo, Carrot2, Vivisimo are some clustering engines. In fig 3, Carrot2 clustering engine gives a clustered result for the query “concord”. By “concord”, a user may mean a town in Massachusetts, USA or a town in New Hampshire or a record label etc. Carrot2 presents users with a list of categories pertaining to word concord. From there, a user can select the desired category. Moreover the clustered view gives the whole picture of the search result and the user can easily understand his/her actual information need.In RF, the performance of the retrieval largely depends on the query structure. The user assessment about document relevance is not always right. The popularity of clustering search result is increasing. [17] Presented a detailed insight into the presently available clustering engines.Clustering has an edge over Relevance feedback as it is simple yet powerful technique that clusters documents without user assessment or reformulating the query.Even a user, who does not know how to fire an idealquery, gets a whole and detailed picture of what he/she actually needs. Many main stream search engines provide

IV. CONCLUSION

Clearly, Clustering is being widely used in web search due to its simplicity. Although it is an overhead computation, still it gives better results. This extra computation time is a onetime deal. Users do not like to wait for response. They want results quickly. With relevance feedback, which is an iterative process, it seems hard to provide results instantly. Relevance feedback systems require prior knowledge of relevance of documents that are used in feedback whereas there is no such requirement in web document clustering. Although both techniques try to improve the retrieval performance, there is a need to measure their performances. As our future work, we would try to evaluate their performance on some actual datasets.

REFERENCES

[1] Salton G. and Buckley C. (1990): ‘Improving retrieval performance by relevance feedback’. Journal of the American Society for Information Science. 41. 4. pp 288-297.

[2] Croft W. and Harper D. (1979).‘Using probabilistic models of information retrieval without relevance information’. Journal of Documentation. 35. 4. pp 285-295.

[3] Rocchio J.J. (1971). ‘Relevance feedback in information retrieval The SMART retrieval system - experiments in automatic document processing. Chapter 14. pp 313-323.

[4] Salton (ed) G. (1971). ‘The SMART retrieval system - experiments in automatic document processing’.

[5] Ruthven, I. and Lalmas, M. (2003). ‘A survey on the use of relevance feedback for information access systems’. Knowledge Engineering Review, 18(2), pp. 95-145.

[6] Ma Y, Lin H (2014) ‘A Multiple Relevance Feedback Strategy with Positive and Negative Models’. PLoS ONE 9(8): e104707. doi:10.1371/journal.pone.0104707.

[7] Zhoua D., Truranb, M., Liua J., Zhanga S. (2013) ‘Collaborative pseudo-relevance feedback’, Expert Systems with Applications, Volume 40, Issue 17, 1 December 2013, Pages 6805–6812.

[8] Steinbach M., Karypis G. and Kumar V. (2000). ‘A Comparison of Document Clustering Techniques’. KDD Workshop on Text Mining.

[9] Zamir O. andEtzioni O. (1998). ‘Web Document Clustering: A Feasibility Demonstration’. Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 46-54.

[10] Alam M. and Sadaf K. (2012). ‘A Review on Clustering of Web Search Result’. Advances in Intelligent Systems and Computing (Springer-Verlag Berlin Heidelberg). Volume 177, pp. 153-159.

[11] Osinski S. (2005). ‘A Concept-Driven Algorithm for Clustering Search Results’. Intelligent Systems, IEEE, Vol. 20, No. 3, pp. 48-54.

[12] Mecca G., Raunich S. and Pappalardo A., (2007). ‘A New Algorithm for Clustering Search Result’, Journal of Data & Knowledge Engineering Volume 62 Issue 3.

[13] Zhang J., Ong K. and Lee V.C.S. (2012). ‘Why Web-Based Pseudo Relevance Feedback Systems Fail’. Kicss, pp. 216-222.

[14] Cao, G., Nie, J.Y., Gao, J., and Robertson, S. (2008). ‘Selecting good expansion terms for pseudo-relevance

feedback'. In SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp 243-250.

[15] An Introduction to Information retrieval, (2009). Cambridge University Press.

[16] Spink, A., Bernard J. J., and H. Cenk O., (2000). Use of query reformulation and relevance feedback by Excite users. Internet Research: Electronic Networking Applications and Policy 10(4), pp 317–328.

[17] Carpenito, C., Osinski S., Romano G. and Weiss D., (2009). 'A Survey of Web Clustering Engines. ACM Computing Surveys,Vol.41,No.3,Article 17.