SEMINAR
HOT TOPICS IN INFORMATIONAL RETREIVAL WS2010/20211

A Report on

Enhancing Cluster Labelling Using Wikipedia

by

David Carmel, Haggai Roitman, Naama Zwerdling

Abdur Raafiu Mohamed Farook
Department of Computer and Communication Technology
Saarland University
abdur.raafiu@gmail.com

Cosmina Croitoru
IMPRS-CS
Department of Computer Science
Saarland University
cosmina@mpi-inf.mpg.de

SAARLAND UNIVERSITY
DATABASES AND INFORMATION SYSTEMS
MAX PLANCK INSTITUT INFORMATIK

UNIVERSITÄT DES SAARLANDES

1 Introduction

Document clustering is used to organize massive amount of textual data in manageable forms such the documents within a cluster are to be chosen as similar as possible and documents in different clusters should be dissimilar. For example, Web search result clustering is usually performed in three main steps:

i. Given a query q, a search engine (e.g., Yahoo!) is used to retrieve a list of results R = (r_1, . . . , r_n);
ii. A clustering C = (C_0,C_1, . . . ,C_m) of the results in R is obtained by means of a clustering algorithm;
iii. The clusters in C are labeled with an appropriate algorithm for visualization and navigational purposes.

Perhaps the most popular application of document clustering is the Google News2 service, which uses document clustering techniques to group news articles from multiple news sources to provide a combined overview of news around the Web.

A lot of research is being done on clustering algorithms and their applications in information retrieval and data mining (step 2, above). Since, comparatively little work has been done on cluster labeling., the paper under review ([1]), has as a main task to devise a high quality clusters labeling algorithm. Traditionally [2], statistical techniques were used to extract labels from the cluster itself. In the present paper, authors introduce *cluster labeling enhancement* by utilizing Wikipedia, the free on-line encyclopedia. They describe a general framework for cluster labeling that extracts candidate labels from Wikipedia in addition to important terms that are extracted directly from the text. The "labeling quality" of each candidate is then evaluated by several independent judges and the top evaluated candidates are recommended for labeling. The experiments done over the resulting cluster

labeling system show that for more than 85% of the clusters in the test collection, the manual label appears in the top five labels recommended by the system.

2 Some background information

Important terms in the cluster content are used to label the cluster that will characterize the cluster in contrast to other clusters. Important terms can be identified by selecting the most frequent terms in the cluster, by extracting the top weighted terms in the cluster centroid, or using any other statistical feature selection techniques [2]. However, a list of significant keywords, or even phrases, will many times fail to provide a meaningful readable label for a set of documents. In many cases, the suggested terms, even when related to each other, tend to represent different aspects of the topic underlying the cluster. In other cases, a good label may not occur directly in the text. Hence user intervention is required to infer a proper label from the suggested terms to successfully describe the cluster's topic. Table 1 bellow shows on the second column the top five important terms extracted for six Open Directory Project (ODP) [4] topics using the JSD selection method. The last column of Table 1 shows that using Wikipedia, labels extracted agree much more with the given human annotated labels (underlined terms).

ODP Category	Top-5 JSD important terms	Top-5 Labels Using Wikipedia Enhancement
Bowling	*bowl*, bowler, lane, bowl center, league	Bowls, *Bowling*, Bowling (cricket), Bowling organisations, Bowling competitions
Buddhism	buddhist, *buddhism*, buddha, zen, dharma	*Buddhism*, History of Buddhism, Buddhism by country, Tibetan Buddhism, Buddhists
Ice Hockey	hockey, nhl, hockey league, coach, head coach	*Ice hockey*, Ice hockey leagues, Hockey prospects, Canadian ice hockey coaches, National Hockey League
Electronics	voltage, high voltage, circuit, laser, power supply	*Electronics*, Power electronics, Diodes, Power supplies, Electronics terms
Tennis Players	wimbledon, tennis, defeat, match today, wta	*Tennis Players*, Tennis terminology, Tennis tournaments, 2002 in tennis, 2000 in tennis
Christianity	church, catholic, ministry, christ, grace	*Christianity*, Christian denominations, Non-denominational Christianity, Christian theology, Christianity in Singapore

Table 1: Top-5 term extracted using JSD and Wikipedia

It follows that the proposed method was far better than using normal statistical method. We note however that, while users may disagree on the exact label of a topic, every user can exploit labeling information for navigation purposes as long as labeling is of high quality.

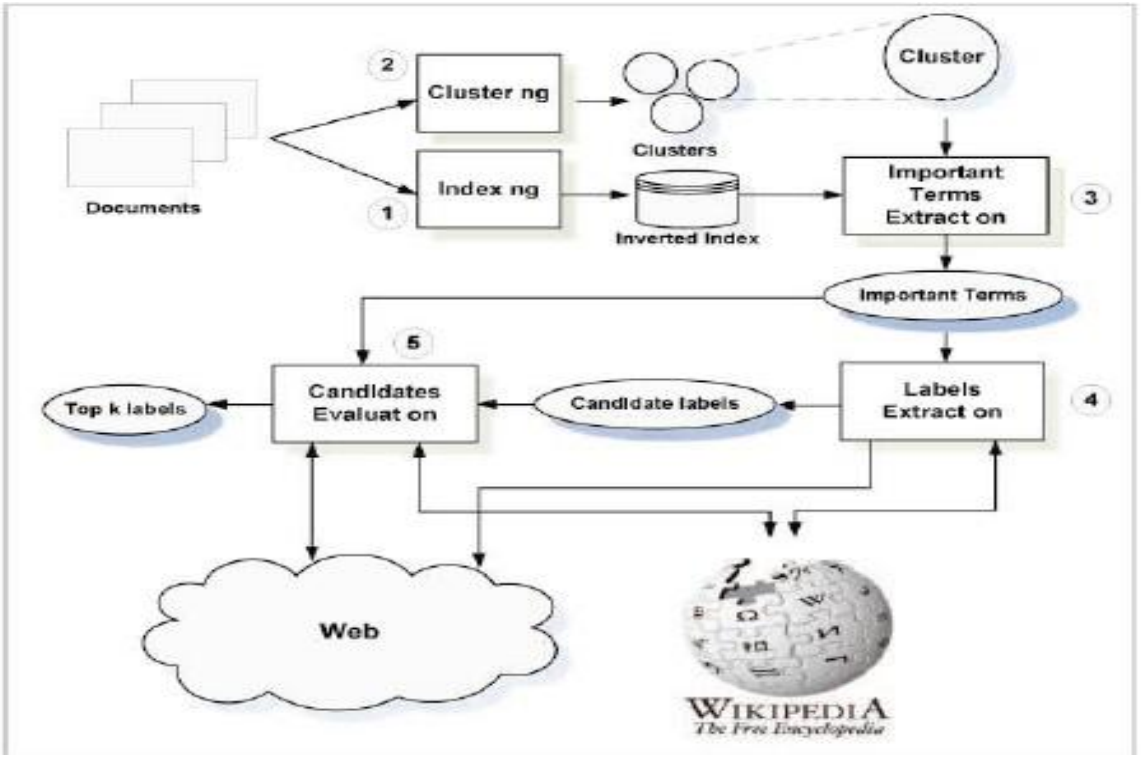
3 Approach used in this paper

The approach they proposed in the paper is as follows, First the system extracts the important terms from the document by using some statistical techniques (e.g., JSD). Second, based on the most important terms extracted, the system forms a query against Wikipedia to find the relevant pages from it. Then from that Wikipedia pages the system extracts the title and categories (i.e., meta data) and these meta data were assigned as candidates. These candidates together to those obtained by the statistical processing, in the first step, are evaluated by several independent judges and the top ranked candidates are selected for labeling the cluster.

The advantage of this approach is to use an external knowledge base for cluster labeling. Even though some similar methods were used in some other papers[3], the novelty is to use the metadata of the Wikipedia pages. The idea of using Wikipedia is motivated since it provides entries on a vast number of "named entities" and very "specialized concepts". In paper [5] ,

entitled *WikiRelate!*, the authors argued that "Including Wikipedia improves the performance of an NLP application processing".

4 General Framework used in the system



The general framework of the system consists of five important modules, namely, indexing, clustering, important terms extraction, labels extraction and candidate evaluation. The system receives a set of textual documents as input. Initially, the documents are parsed and indexed and an inverted index is generated. This index is primarily used by other components for gathering term statistics. The documents are then clustered using the clustering component. For each generated cluster, the system extracts a set of important terms that are estimated to best represent the content of the documents of the cluster using JSD.

JSD, Jensen-Shannon divergence, measures distances between sets of documents and sets of terms. It is a symmetric version of the Kullback-

Leibler divergence. More precisely, if *P(w)* and *Q(w)* are distributions over the words in the collection wεW, JSD is defined as $DJ(P|Q)=\sum_{w\varepsilon W}P(w)\log (P(w)/M(w))+ \sum_{w\varepsilon W}Q(w)\log (Q(w)/M(w))$, where $M(w)=1/2(P(w)+Q(w))$. JSD is not a distance, but its square root is. JSD is preferred over other distance measures such as cosine distance, since when measuring distances between documents and sets of terms, the collection statistics can be naturally incorporated into measurements[6]. Each term is scored according to its contribution to the JSD distance between the cluster and the collection. The cluster important terms are then used to identify a list of candidate labels for the cluster. Candidate labels can be selected from the set of important terms or from external resources (e.g., Wikipedia, or the general web).
Candidate labels are evaluated by several judge systems. Each judge gets as an input the set of candidate labels and the set of the cluster's important terms. Then, each judge evaluates the candidates according to its evaluation policy. The scores of all judges are then aggregated and the labels with the highest aggregated scores are returned. The judges used are two instantiations of **Mutual Information** (MI) (which scores each candidate by the average pointwise mutual information (PMI) of the label with the set of the cluster's important terms, with respect to a given external textual corpus) and two instantiations of **Score Propagation** (SP) (which scores each candidate label with respect to the scores of the documents in the result set associated with that label).

5 Discussion about experiments

Two data collections were considered for the experiment purpose namely 20NG(news group) and ODP(open directory project). 20,000 and 10,000 documents were used for the experiment, respectively. From the first

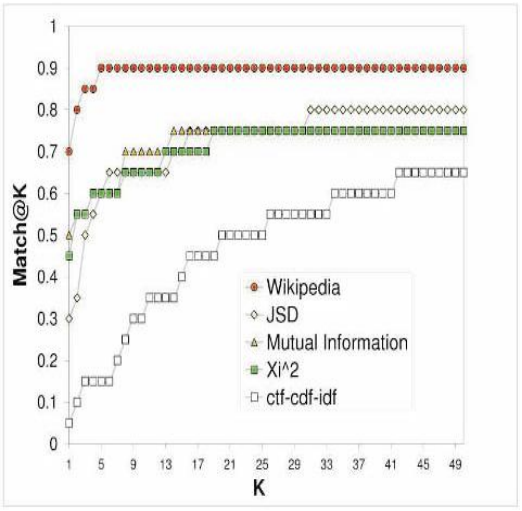
collection, 20 categories have been selected and from each category 1000 documents were included. The second collection selected 100 categories and 1000 documents per category. This kind of selection of the data collection provides an implicitly set of clusters, which were already manually labeled. For the given collection of clusters the system returns up to k-labels for each cluster. The system yields very good results, but as we are observe, this is not quite conformal with the general framework described above (it assumes a very artificial setting of the clustering and indexing modules).

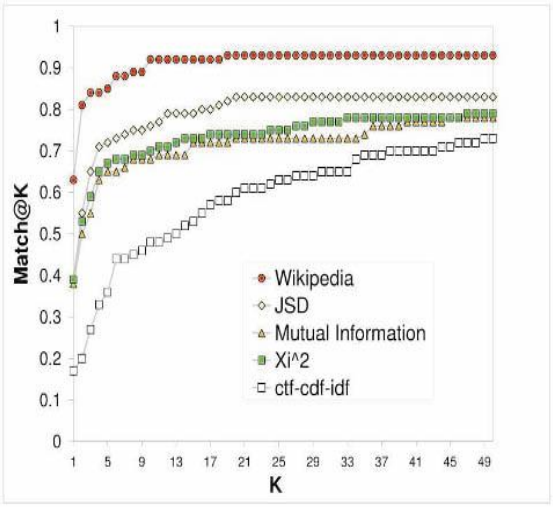
6 Discussions about evaluation

It is considered that the system was working on lower bounds, since all the documents were manually labeled documents. So the evaluation system may expect the output of the system should be similar or identical to the manual labels. They used two methods to evaluate the system in each and every phase,

- **Match@K**
- **Mean Reciprocal Rank (MRR@K)**

Firstly, it is evaluated the effectiveness of using Wikipedia in enhancing the cluster labeling . Four different feature selection methods, with and without Wikipedia, were compared and the output of the graph describes that using Wikipedia yields more relevant labels for the clusters. Also, JSD technique, used to extract the important terms, scores a higher value when compared to other feature selection methods.

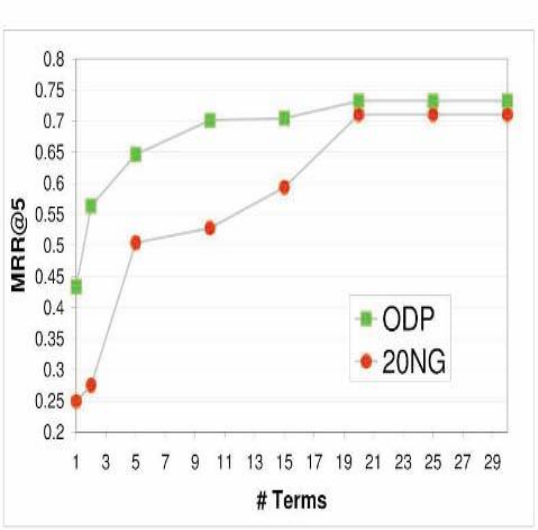


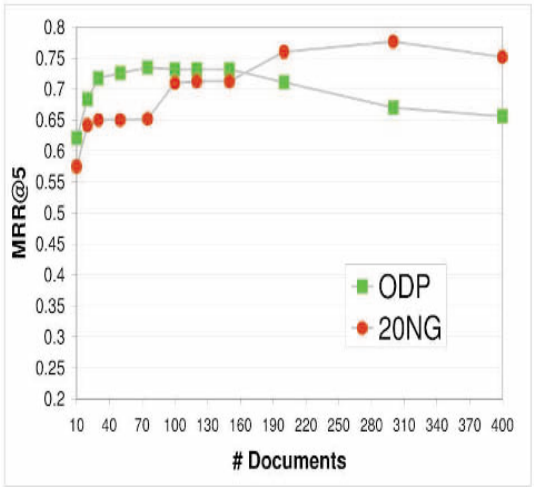


There are two significant parameters that can affect the quality of Wikipedia's labels:

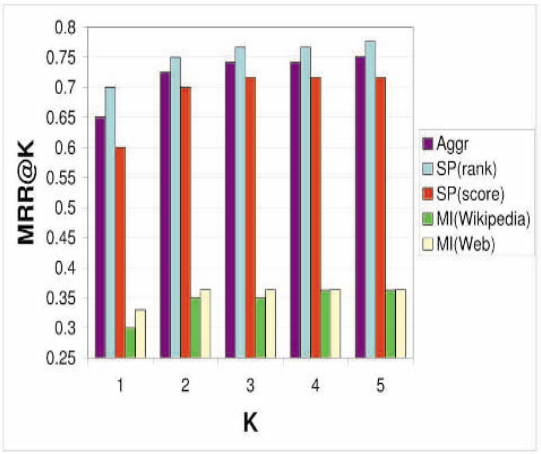
- The number of important terms that are used to query Wikipedia
- The number of top scored results from which candidate labels are extracted

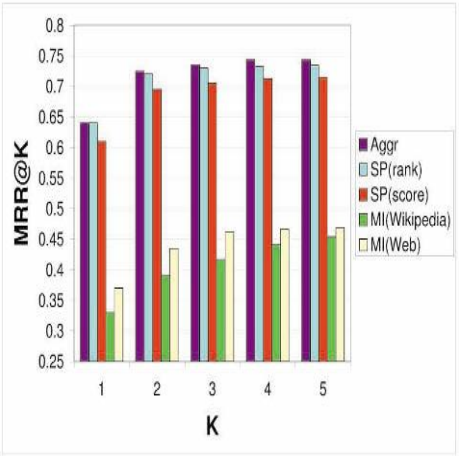
As expected, the result of this evaluation part shows that when there is a increment in number of query terms there will be plateau and when in the case of number of documents the system performance degrades.



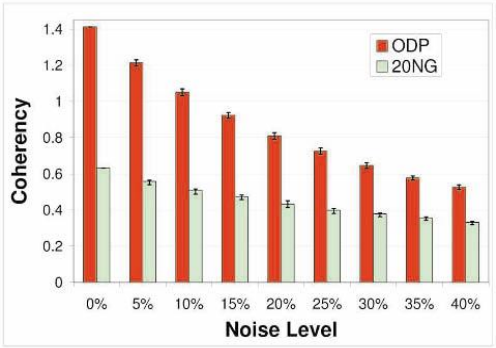


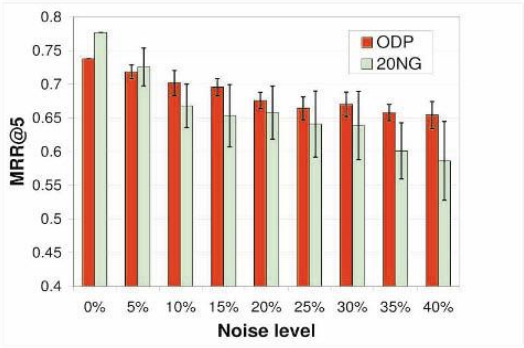
The evaluation of judges was done by their ability to identify the correct labels. From the evaluation results overall, among the four different judges, the SP(rank) judge performs the best





Finally, the effect of cluster coherency is evaluated, since when there is no proper coherencies in the cluster the system won't yield good results. The system performance is evaluated by adding noise into the clusters, i.e., by making each document in one cluster to swap with document in other cluster with probability *p*. This means that, introducing more noise, we obtain a less coherent clusters and therefore, the MRR score drops. Nevertheless, the drop in MRR score per noise level is quite moderate for both datasets which implies that the proposed system is robust and has good resiliency to noise:





7 Final comments

The paper under review showed that using Wikipedia as a thesaurus and applying this thesaurus as post-processing step to statistical labeling approaches improves cluster labeling dramatically.

However, extraction from Wikipedia achieve high precision and recall on well-populated classes of articles, they fail in a larger number of cases, largely because incomplete articles and infrequent use of infoboxes.

Arguably not all the Wikipedia articles are of equally high quality. Using the meta data of such articles could introduce large noise that could influence the final ranking of the terms. Perhaps, the use of some weights associated to the "quality" or "trust" of information enhanced could be more accurate.

The Wikipedia category tree is an example of a folksonomy, namely a collaborative tagging system that enables the users to categorize the content of the encyclopedic entries. Folksonomies as such do not strive for correct conceptualization in contrast to systematically engineered ontologies. In the same time, WordNet [7] represents a well structured taxonomy organized in a meaningful way. It would be interesting if the labeling system is more accurate by using combination of WordNet and Wikipedia (not only in the evaluation phase).

Especially Wikipedia shows the biggest dependency on hierarchical information, followed by the ODP dataset. Intuitively, the hierarchical structure should influence the labeling accuracy. As noted in [8], the impact of hierarchical structures on the labeling accuracy is yet unclear. A weak point of the paper is that it is not concerned with hierarchical structures.

Last, but not least, we note that the novel idea used in this paper to improve label quality by using the meta data gathered from Wikipedia could be viewed as an interesting argument for the Tim Berners-Lee principle "Create Knowledge out of Interlinked Data" [9].

8 References

[1]. David Carmel, Haggai Roitman, Naama Zwerdling Enhancing Cluster Labeling using Wikipedia. SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.

[2]. C. D. Manning, P. Raghavan, and H. Schutze. Introduction to Information Retrieval. Cambridge University Press,2008.

[3]. Z. S. Syed, T. Finin, and A. Joshi. Wikipedia as an ontology for describing documents. In ICWSM '08, 2008.

[4]. Open Directory Project (ODP). http://www.dmoz.org/.

[5]. Michael Strube, Simone Paolo Ponzetto: WikiRelate! Computing Semantic Relatedness Using Wikipedia. AAAI 2006

[6]. David Carmel, Elad Yom-Tov, Adam Darlow, Dan Pelleg: What makes a query difficult? SIGIR 2006: 390-397

[7]. G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross, K. Miller. WordNet: An online lexical database. Int. J. Lexicograph. 3, 4, pp. 235-244, 1990.

[8]. Markus Muhr, Roman Kern, Michael Granitzer: Analysis of structural relationships for hierarchical cluster labeling. SIGIR 2010: 178-185

[9]. Tim Berners-Lee, http://www.w3.org/DesignIssues/LinkedData.html