Multilingual Document Clustering using Wikipedia as External Knowledge

Kiran Kumar N, Santosh GSK, Vasudeva Varma

International Institute of Information Technology, Hyderabad, India
{kirankumar.n, santosh.gsk}@research.iiit.ac.in, vv@iiit.ac.in

Abstract. This paper presents Multilingual Document Clustering (MDC) on comparable corpora. Wikipedia, a structured multilingual knowledge base, has been highly exploited in many monolingual clustering approaches and also in comparing multilingual corpora. But there is no prior work which studied the impact of Wikipedia on MDC. Here, we have made an in-depth study on availing Wikipedia in enhancing MDC performance. We tried to utilize its knowledge structure (Crosslingual links, Category, Outlinks, Infobox information, etc.) to enrich the document representation for clustering multilingual documents. By avoiding language-specific tools, this approach has become a general framework which can be easily extensible to other languages. We have experimented with bisecting k-means clustering algorithm on a standard dataset provided by FIRE[1] for their 2010 Adhoc Cross-Lingual document retrieval task on Indian languages. We have considered English and Hindi datasets. The system is evaluated using F-score and Purity measures and the results obtained are encouraging.

Keywords
Multilingual Document Clustering, Wikipedia, Document Representation

1 Introduction

The amount of information available on the web is steeply increasing with globalization and rapid development of internet technology. Due to various contributers across the world, the information is present in varied languages. The increasing amount of documents written in different languages, creates a need to develop applications to manage that massive amount of varied information. MDC proves to be very useful in processing and managing multilingual information present on the web. MDC involves dividing a set of n documents written in different languages, into various clusters so that the documents that are semantically more related belong to the same cluster. It has got applications in various streams such as Cross Lingual Information Retrieval (CLIR) [1], training of the parameters in statistical machine translation, or the alignment of parallel and non parallel corpora, among others.

[1] Forum for Information Retrieval Evaluation - http://www.isical.ac.in/∼clia/

In traditional text clustering methods, documents are represented as "bag of words" without considering the semantic information of each document. For instance, if two documents use different collections of keywords to represent the same topic, they may be falsely assigned to different clusters. This problem arises due to the lack of shared keywords, although the keywords they use are probably synonyms or semantically associated in other forms. The most common way to solve this problem is to enrich the document representation with an external knowledge or an ontology. Wikipedia is one such multilingual knowledge base that supports 257 language editions as of now. In this paper, we have made an in-depth study of different ways of exploiting its huge multilingual knowledge base in enhancing the performance of MDC. Using Wikipedia over other knowledge resources or ontologies has got advantages like:

1. Wikipedia has the multilingual content along with its metadata at a comparable level. By availing the cross-lingual links this information can be aligned.
2. Wikipedia supports 257 active language editions. As Wikipedia acts as a conceptual interlingua with its cross lingual links, our approach is scalable to other languages with relative ease.
3. With its wide access to many editors, any first story or hot topic gets updated in Wikipedia which can enhance the clustering performance of news related documents. Hence, this approach also addresses the future growth of multilingual information.

The rest of the paper is organized as follows: Section 2 talks about the related work. Section 3 describes our clustering approach in detail. Section 4 presents the experiments that support our approach. Finally we conclude our paper and present the future work in Section 5.

2 Related Work

MDC is normally applied on parallel [2] or comparable corpus [3–5]. In the case of the comparable corpora, the documents usually are news articles. Considering the approaches based on translation technique, two different strategies are employed:

1. Translate the whole document to an anchor language.
2. Translate certain features of the document, that best describes it, to an anchor language.

The work proposed in [6] uses bilingual dictionaries for translating Japanese and Russian documents to English (anchor language). Translating an entire document into an anchor language is often not preferred due to the time overhead involved. When the solution involves translating only some features, first it is necessary to select these features (usually entities, verbs, nouns) and then translate them using a bilingual dictionary or by consulting a parallel corpus. Montalvo *et*

al. [7] presented a MDC approach using only cognate named entities extracted from the multilingual documents.

Limited work has been done where existing knowledge structures, like EUROVOC thesaurus [8] etc., were used for measuring cross-lingual document similarity. However, the EUROVOC has support only for European languages. Pouliquen *et al.* [9] proposed a method to extract language-independent text features using gazetteers and regular expressions besides thesaurus and classification systems. However, the gazetteers support only a limited set of languages. Such resources doesn't satisfy the need to deal with information in diverse languages. Hu *et al.* [10] exploited Wikipedia knowledge base in the realm of monolingual document clustering. But none of the revised works used Wikipedia to enhance the performance of MDC. We explain our MDC approach which is based on efficient exploitation of Wikipedia in Section 3.

3 Proposed Approach

In this section, we detail the various phases involved in the proposed approach for clustering multilingual documents. The pool of English and Hindi text documents are first represented as (basic) document vectors. These document vectors are enriched with Wikipedia knowledge base to obtain additional vectors. The basic document vectors along with enriched document vectors are linearly combined for measuring the document similarity. Based on the similarity measure, clusters are formed, separately for English and Hindi documents. Similarity between the centroids are measured to combine these clusters, details are explained in a later section. The results showcases the effectiveness of the proposed approach. The further sections details every phase of our approach.

3.1 Document Representation

All the English and Hindi text documents are represented in the classical vector space model [11]. It represents the documents as a vector of keyword-based features following 'bag of words' notation having no ordering information. The values of the vector are TFIDF scores. Instead of maintaing a stopword list for every language, any word that appears in more than 50% of the documents is considered a stopword. Besides removing stopwords, the document vectors still contain certain umwanted words that lead to noise. So, we have considered only the top-k keywords in each document, based on their TFIDF scores. We have experimented by considering k values from 40% to 100% with an increment of 10%. Best cluster results are achieved for k=50%.

3.2 Enriching the Document Representation

To characterize the every growing content and coverage of Wikipedia, its articles are annotated and categorized. The content of a Wikipedia article is annotated

by hyperlinks (references) to other articles and they denote the "Outlinks" for that article. Such Outlinks create interlinks between articles. Every article is a description about a single topic or a concept. Equivalent topics (Concepts) that are represented with different phrases are grouped together by redirected links. Meanwhile, it contains a hierarchical categorization system, in which each article belongs to at least one Category. In regard to all the above features, Wikipedia has become a potential resource which can be exploited in enhancing text document clustering.

As it was said earlier that the BOW model doesn't consider the semantics of the words, so two documents with different collections of keywords representing the same topic, may be falsely assigned to different clusters. Hu *et al.* [10] have presented an approach in which they have overcome this problem by enriching their document representation by incorporating Wikipedia Concepts vector and Categories vector for monolingual document clustering.

These document vectors form the baseline (basic) keyword vector for our experiments. Along with the Wikipedia Concepts and Categories, in this paper

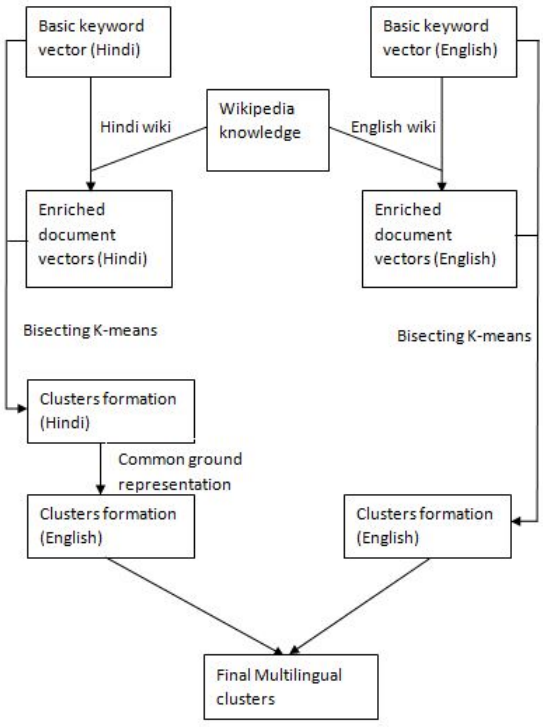

Fig. 1. MDC Approach

we have also explored the potential Cross-lingual links, Outlinks and Infobox information. The Infobox of an article contains certain statistical information which represents the most prevalent information need in that article. Hence it can be helpful in deciding the topic of a document and thereby improving the cluster quality.

Wikipedia databases are preprocessed and the title of every Wikipedia page is mapped to its corresponding Categories, Outlinks, Infobox and the Synonyms with the help of lucene indexer. This mapping process is divided into three steps:

1. For every page of Wikipedia, its title is extracted along with its corresponding Outlinks, Categories and Infobox information.
2. All the redirections are tracked, as they are considered to be the synonyms for Wikipedia titles.
3. An inverted index is built using Lucene indexer[2]. Lucene builts an index with key-value pairs. Every Wikipedia title forms a pair with each of its corresponding Categories, Outlinks, Infobox and Synonyms. All such pairs are indexed and stored.

A separate index is created for English and Hindi Wikipedia databases.

For every document, its basic keyword vector is used for obtaining three enriched vectors from Wikipedia structure namely Outlink vector, Category vector and Infobox vector. For every term in the the basic keyword vector, its corresponding match for a Wikipedia title is fetched. If present, the Outlink, Category and Infobox terms from the corresponding Wikipedia article and also from its synonym (redirection) articles are extracted. These terms add up to the Outlink vector, Category vector and Infobox vector of that document respectively. This is repeated for all the terms in the keyword vector. The values are the TFIDF scores of those terms.

Addition of all possible terms would lead to noise in each of these additional vectors. So, we had to considered only the top-k terms in each of these additional vectors in the similar way as we did for the basic keyword vector. The main advantage of our approach over existing approaches is that we are adding semantic information only to a subset of document terms that are considered to be important owing to the fact that there might be unimportant words or outliers for which the addition of extra information might lead to distortion of the original clustering.
E.g. Consider the following document *"It was an exciting match between India and Pakistan. . . . . Pakistan prime minister Parvez Musharraf awarded the man of the match to Sachin Tendulkar. . . ."* .
This document is actually about a cricket match. If we enrich the entire document representation and add unwanted information about Parvez Musharraf, prime minister, etc., it degrades the clustering performance. Considering only the top-k keywords has also reduced the computation time.

[2] http://lucene.apache.org/java/docs/index.html

3.3 Document Clustering

Document clustering is an automatic grouping of text documents into clusters so that documents within a cluster have high similarity in comparison to one another, but are dissimilar to documents in other clusters. Various clustering approaches (such as Hierarchical clustering, Partitioning clustering, etc.) are available to cluster the enriched documents. We used Bisecting k-means algorithm [12] for the clusters formation as it combines the strengths of partitional and hierarchical clustering methods by iteratively splitting the biggest cluster using the basic k-means algorithm. For the evaluation of bisecting k-means algorithm, we have experimented with fifteen random k values between 30-70 and the average is considered as the final clustering result.

Steinbach *et al.* [12] compared different algorithms and concluded that bisecting k-means performs better than the standard k-means and agglomerative hierarchical clustering. Basic k-means is a partitional clustering algorithm based on the basic vector space model. Bisecting k-means algorithm is applied on the enriched documents vectors along with the basic keyword vector to obtain separate set of clusters for English and Hindi documents respectively. At the heart of a clustering algorithm is a similarity measure. We choose the cosine distance, which measures the similarity of two documents by calculating the cosine of the angle between them. Each document is represented with four vectors: basic Keyword vector, Category vector, Outlink vector and Infobox vector.

The similarity between two documents d_i and d_j is defined as follows:

$$sim(d_i, d_j) = sim^{basic\_keyword} + \alpha * sim^{Category} + \beta * sim^{Outlink} + \gamma * sim^{Infobox} \quad (1)$$

Here, sim(d_i,d_j) gives the cosine similarity of the documents d_i,d_j. The sim is calculated as:

$$sim = cos(v_i, v_j) = (v_i.v_j)/(|v_i| * |v_j|) \quad (2)$$

where v_i and $v_j \in$ {basic keyword, Category, Outlink, Infobox} vectors of documents d_i and d_j respectively. The coefficients α, β and γ indicates the importance of Category vector, Outlink vector, and Infobox vector in measuring the similarity between two documents. A similarity measure similar to Eq.(1) was proposed by Hu *et al.* [10] where the Wikipedia Concepts and Categories were used for clustering the monolingual documents. The Wikipedia Category information was also utilized by Hu *et al.* [13], Wang and Domeniconi [14] for monolingual text clustering and classification respectively.

To compare multilingual clusters, the document vectors are mapped onto a common ground representation (English), the details are explained in the next section. For a cluster in the Hindi set of clusters, its centroid is calculated by taking the average of all document vectors of that cluster. Given a document in a cluster, its representing vector is a linear combination of the basic keyword vector and the enriched vectors. Experiments are conducted by considering many such

possible linear combinations. The similarities of a Hindi cluster centroid with the centroids of each of the English clusters are measured. The calculated values are noted in a sorted order. This is repeated for all the Hindi clusters. The two clusters (one English, one Hindi) with the highest similarity value are merged to form a multilingual cluster. This step is repeated with the remaining set of clusters and finally multilingual clusters are achieved.

3.4 Common Ground Representation

In order to calculate similarity between two multilingual documents we may have to use language resources (dictionaries etc.) or language specific tools (lemmatizers, NER, POS tagger, etc.). With the inclusion of any language specific tool, it takes painful development process in order to reimplement an approach for a language with fewer resources, which very much limits the extensibility of an approach. So we avoided such issues by eliminating the usage of any language specific tools in our approach. Instead we have used structured Wikipedia multilingual content and a bilingual dictionary. To cover a broader set of terms, we preferred to use the Shabdanjali dictionary[3].

Proper nouns play a pivotal role in determing the theme of a document that can greatly enhance similarity calculation. Dictionaries, in general, doesn't cover many proper nouns. Transliteration might be highly helpful in identifying the proper nouns, but it requires parallel transliterated (English-to-Hindi) word lists to build even a language-independent statistical transliteration technique [15]. Acquiring such word lists is a hard task when one of the language is a minority language. Including transliteration comes at the cost of reducing the extensibility of our approach.

As an alternative, we utilized the cross-lingual links that exist in Wikipedia multilingual databases (English and Hindi). The cross lingual links interrelates Wikipedia articles of different languages. All these articles follow the constraint of sharing identical topics and their titles are verified to be aligned. We have availed these alignments in creating a Wiki dictionary to handle proper nouns. This method is language-independent and is easily scalable for other languages. We have mapped all the Hindi documents vectors onto English using bilingual dictionary and the Wiki dictionary.

Modified Levenshtein Edit Distance Measure: In all the similarities calculated above, the terms are compared using the Modified Levenshtein edit distance as a string distance measure. In many languages, words appear in several inflected forms. For example, in English, the verb 'to walk' may appear as 'walk', 'walked', 'walks', 'walking'. The base form, 'walk', that one might look up in a dictionary, is called the lemma for the word. The terms are usually lemmatized to match the base form of that term. Lemmatizers are available for English and

[3] http://ltrc.iiit.ac.in/onlineServices/Dictionaries/Dict_Frame.html

many other European languages. But the lemmatizers support is very limited in the context of Indian Languages. So, we have modified the levenshtein edit distance metric to replace the purpose of lemmatizers by adding certain language-independent rules. Henceforth, it can be applied for any language. This modified levenshtein edit distance would help us in matching a word in its inflected form with its base form or other inflected forms. The rules are very intuitive and are based on three aspects:

1. Minimum length of the two words
2. Actual Levenshtein distance between the words
3. Length of subset string match, starting from first letter.

No language specific tools are used in this approach. However, to ensure the accuracy, we have worked out alternatives (like Wiki dictionary, Modified Levenshtein Edit distance, etc.) which are language-independent.

4 Experimental Evaluation

We have conducted experiments using the FIRE 2010 dataset available for the ad-hoc cross lingual document retrieval task. The data consists of news articles collected from 2004 to 2007 for each of the English, Hindi, Bengali and Marathi languages from regional news sources. There are 50 query topics which are equivalently represented in each of these languages. We have considered the English and Hindi articles for our experiments. We used the topic-annotated documents in English and Hindi to build clusters. To introduce noise, we have added non-topic documents. The noisy considered constitue 10 percent of the number of topic documents. Some topics are represented by 8 or 9 documents whereas others are represented by about 50 documents. There are 1563 documents in the resulting collection of 50 topics, 650 are in English and 913 in Hindi. Cluster quality is evaluated by F-score [12] and Purity [16] measures. F-score combines the information of precision and recall. To compute purity, each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned documents and dividing by total number of documents.

4.1 Wikipedia Data

Wikipedia releases periodic dumps of its data for different languages. We used the latest dump consisting of 2 million English documents and 50,000 Hindi documents. The data was present in XML format. The required Wikipedia information such as Categories, Outlinks, Infobox and Redirections are extracted and processed for the formation of vectors.

4.2 Discussion

In our experiments, clustering based on keyword vector is considered as the baseline. Various linear combinations of Keyword, Category, Outlink and Infobox

Table 1. Clustering schemes based on different combinations of vectors

Notation	**F-Score**	**Purity**
Keyword (baseline)	0.532	0.657
Keyword_Category	0.563	0.672
Keyword_Outlinks	**0.572**	0.679
Keyword_Infobox	0.544	0.661
Category_Outlinks	0.351	0.434
Category_Infobox	0.243	0.380
Outlinks_Infobox	0.248	0.405
Keyword_Category_Outlinks	0.567	**0.683**
Keyword_Outlinks_Infobox	0.570	0.678
Keyword_Category_Infobox	0.551	0.665
Category_Outlinks_Infobox	0.312	0.443
Keyword_Category_Outlinks_Infobox	0.569	0.682

vectors are examined in forming clusters. The cluster quality is determined by F-score and purity measures. Table 1 displays the results obtained. As mentioned earlier, the coefficients α, β and γ determine the importance of Category, Outlink and Infobox vectors respectively in measuring the similarity between two documents. To determine the α value, we have considered Keyword and Category vectors to form clusters. Using Eq.(1) experiments are done by variying the α values from 0.0 to 1.0 by 0.1 increment (β and γ are set to 0). The value for which best cluster result is obtained is set as the α value. Similar experiments are repeated to determine β and γ values. In our experiments, it was found that setting $\alpha = 0.1$, $\beta = 0.1$ and $\gamma = 0.3$ yielded good results.

From Table 1, it can be observed that our experiments using external knowledge resource has performed better than the baseline. Moreover, the Outlinks information has proved to perform better than Categories followed by Infobox information. Outlinks are the significant informative words in a Wikipedia article which has references (hyperlinks) to other articles. As the Outlinks nearly overlap the context of a document, that might have improved the results better than the rest. With the Categories, we get generalizations of a Wikipedia topic. Considering the Categories, the documents are compared at an abstract level which might have declined the results compared to Outlinks. The Infobox provides vital statistical information of a Wikipedia article. However, its information is inconsistent across all articles which addresses its poor performance when compared with others.

5 Conclusion and Future work

In this paper, we proposed an approach for enhancing MDC performance by exploiting different dimensions (Cross-lingual links, Outlink, Category and Infobox

information) of Wikipedia and tested it with a bisecting k-means clustering algorithm. Our results showcases the effectiveness of Wikipedia external knowledge in enhancing MDC performance. The Outlinks information has proved to be crucial in improving the results, followed by Categories and Infobox information. We have avoided any use of language-specific tools in our approach, instead we have created alternatives (like Wiki dictionary, Modified Levenshtein Edit Distance, etc.) to ensure the accuracy. Provided a bilingual dictionary, this approach is extensible for many other language pairs that are supported by Wikipedia. It is easy to reproduce and furthermore considers the future growth of information across languages.

We plan to extend the proposed approach, which implements only static clustering to handle the dynamic clustering of the multilingual documents. In addition to the aligned titles of Wikipedia articles that have cross lingual links, we further plan to consider the Category and Infobox information in building a robust dictionary, eliminating the use of bilingual dictionary and hence achieving a language independent approach. We would also like to consider comparable corpora of different languages to study the applicability of our approach.

References

1. Pirkola, A., Hedlund, T., Keskustalo, H., Järvelin, K.: Dictionary-based cross-language information retrieval: Problems, methods, and research findings. Information Retrieval **4** (2001) 209–230
2. Silva, J., Mexia, J., Coelho, C., Lopes, G.: A statistical approach for multilingual document clustering and topic extraction form clusters. In: Pliska Studia Mathematica Bulgarica, Seattle, WA (2004) 207–228
3. Rauber, A., Dittenbach, M., Merkl, D.: Towards automatic content-based organization of multilingual digital libraries: An english,french, and german view of the russian information agency novosti news. In: Third All-Russian Conference Digital Libraries: Advanced Methods and Technologies, Digital Collections, Petrozavodsk, RCDI (2001)
4. Leftin, L.J.: News blaster russian-english clustering performance analysis. Technical report, Columbia computer science Technical Reports. (2003)
5. Romaric, B.M., Mathieu, B., Besançon, R., Fluhr, C.: Multilingual document clusters discovery. In: RIAO. (2004) 1–10
6. Evans, D.K., Klavans, J.L., McKeown, K.R.: Columbia news blaster: multilingual news summarization on the web. In: HLT-NAACL–Demonstrations '04: Demonstration Papers at HLT-NAACL 2004, Morristown, NJ, USA, Association for Computational Linguistics (2004) 1–4
7. Montalvo, S., Martínez, R., Casillas, A., Fresno, V.: Multilingual document clustering: an heuristic approach based on cognate named entities. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (ACL), Morristown, NJ, USA, Association for Computational Linguistics (2006) 1145–1152
8. Steinberger, R., Pouliquen, B., Hagman, J.: Cross-lingual document similarity calculation using the multilingual thesaurus eurovoc. In: Proceedings of the Third

International Conference on Computational Linguistics and Intelligent Text Processing (CICLing '02), London, UK, Springer-Verlag (2002) 415–424
9. Pouliquen, B., Steinberger, R., Ignat, C., Käsper, E., Temnikova, I.: Multilingual and cross-lingual news topic tracking. In: COLING '04: Proceedings of the 20th International conference on Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics (2004) 959
10. Hu, X., Zhang, X., Lu, C., Park, E.K., Zhou, X.: Exploiting wikipedia as external knowledge for document clustering. In: KDD '09: Proceedings of the 15th ACM SIGKDD International conference on Knowledge discovery and data mining, New York, NY, USA, ACM (2009) 389–396
11. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM **18** (1975) 613–620
12. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: TextMining Workshop, KDD. (2000)
13. Hu, J., Fang, L., Cao, Y., Zeng, H.J., Li, H., Yang, Q., Chen, Z.: Enhancing text clustering by leveraging wikipedia semantics. In: SIGIR '08: Proceedings of the 31st annual International ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (2008) 179–186
14. Wang, P., Domeniconi, C.: Building semantic kernels for text classification using wikipedia. In: KDD '08: Proceeding of the 14th ACM SIGKDD International conference on Knowledge discovery and data mining, New York, NY, USA, ACM (2008) 713–721
15. Ganesh, S., Harsha, S., Pingali, P., Varma, V.: Statistical transliteration for cross language information retrieval using hmm alignment and crf. In: Proceedings of 3rd International Joint Conference on Natural Language Processing, IJCNLP, Hyderabad, INDIA (2008)
16. Zhao, Y., Karypis, G.: Criterion functions for document clustering: Experiments and analysis. Technical report, Department of Computer Science, University of Minnesota. (2002)