Aspect-Oriented Document Clustering for Facilitating Retrieval Process

M. Hosseinia, K. Badie, and A. Moeini

***Abstract*—Retrieved documents from queries are clustered to help users find information needed more significant in information retrieval. There are some frequent queries try finding information on an issue from the aspect of another issue. But current methods of clustering do not pay attention to the concept of the aspect included in these queries after retrieval process. In this paper we introduce aspect-oriented document clustering to group documents more significant and based on a special point of view. In our approach, text documents are represented based on a special aspect and the similarity between them is computed on the basis of its features. We use Wikipedia as background knowledge to emphasize and enrich the concept of the aspect. Then we evaluate the proposed approach with selected documents from two popular datasets, 20 Newsgroups and Reuters 21578. Results demonstrate that aspect-oriented clustering enhances clustering performance of those documents which can be equivalent to retrieved documents from aspect based queries significantly.**

***Index Terms*—Aspect based representation, aspect-oriented clustering, clustering, Wikipedia.**

I. Introduction

Information is essential to us all times in our lives. But achieving information needed without organizing them is impossible while its sources are growing quickly.

Document clustering is an unsupervised text mining task for organizing information through which documents are grouped into meaningful clusters [1]. It has been applied to group retrieval results from search engines for helping users find information needed more significant. The queries that users usually ask to get information on an issue from the aspect of another issue are frequent and in many of them the concept of aspect is hidden. Current methods of clustering do not pay attention to this concept. By ignoring this concept in clustering of retrieval results, the output is huge amount of documents which just includes issues represented in query with no special consideration on those documents that contain same issues from the view point of other issues.

We propose a new clustering approach to issue this problem, called aspect–oriented document clustering. There is little information available in literature about aspect-orientedness and this is the first time that it is being applied to the issue of clustering in general and document clustering in particular. By aspect–oriented clustering we mean that similarity between documents is evaluated on the basis of a special point of view. So, documents in a same cluster have maximum similarity based on the aspect and are dissimilar with documents in the other clusters. Our motivation to present this approach is to make storage space get close to query space to gather aspect based similar documents in the form of a cluster.

To introduce this approach, an external knowledge for finding and emphasizing features of an aspect in documents is essentially needed. In our approach we make use of Wikipedia [2], the largest multilingual free content encyclopedia on the Internet, as an adequate source for required information. We test our approach on selected documents from two datasets. Evaluation results show significant improvements in clustering performance.

Manuscript received June 19, 2012; revised August 5, 2012.

M. Hosseinia was with the university of Tehran, Tehran, Iran (e-mail: mhosseinia@ut.ac.ir).

K. Badie is with the Research Institute for ICT (ITRC), Tehran, Iran (e-mail: k_badie@itrc.ac.ir).

A. Moeini is with the university of Tehran, Tehran, Iran (e-mail: moeini@ut.ac.ir).

II. Background

Wikipedia is the largest multilingual free content encyclopedia on the Internet launched in 2001 by Jimmy Wales and Larry Sanger with the aim at describing each concept in the world. It is a huge information resource which grows quickly and "Its English language articles alone are 10 times the size of the English Britannica [3], its nearest rival" [4].

Each article, the basic unit of information in Wikipedia, describes a unique concept belongs to several categories and there is a unique article for each concept [4]. *Redirects* link equivalent terms to an article express synonymy. Articles are connected together with internal links exhibits relationships. Due to present these different type of information, leveraging Wikipedia as an external knowledge to improve text mining tasks like document clustering has been increased recently.

An effort for exploiting Wikipedia in document clustering by using Wikipedia concepts, redirects and category information can be found in [5]. They enriched text documents and developed two approaches for mapping text documents to Wikipedia concepts. Their results showed that combining category information with document content generates the best results in most cases. But [6] concluded that representing documents based on concept and document information lead to enhance clustering performance. They also found out that, article titles and its compressed version based document representation has good effect on clustering results due to ignore unusual and unimportant terms. Other works in this area include [7] clustered short text documents with use of Wikipedia data to solve the problem of information overloading in news or blog feeds.

Reference [8] proposed a method to exploit the importance of N-grams in a document and Wikipedia based additional

knowledge for document clustering. Their results show improvement in performance over bag-of-words based representation.

III. The Proposed Approach to Aspect-Oriented Document Clustering

A. Concept of the Aspect

An issue can be described from the aspect of some other different issues. To find the relationship between the interested aspect and the issue, it is essential to emphasize features of the aspect. E.g. when we consider at "*international communications from the aspect of economics,"* the concepts like microeconomics, macroeconomics, international economics, inflation, monetary policy, etc are some features of economics.

B. Aspect-Orientedness in Document Clustering

When we want to mark out a ball based on its corresponding usage, a ball game kind is the considering aspect in marking out a ball, like the right hand one in Fig. 1 for football game.

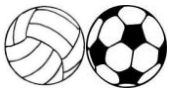

Fig. 1. Marking out a ball.

Now, suppose that the ball is a document collection which we want to group it based on an aspect like the task of marking out a ball. A document collection can be grouped in different ways based on the point of view in clustering. But what do we mean by aspect-oriented document clustering exactly?

E.g. we want to cluster document from the view point of economics. For economics-oriented clustering in vector space model, features of a document vector correspond to features of economics and demonstrate economics-based representation of a document. In this case, two documents:

1) Can be similar while associate with aspect of economics and lie in same cluster after performing economics-oriented clustering.
2) Can be dissimilar while associate with aspect of economics and lie in dissimilar clusters after performing economics-oriented clustering.
3) Do not associate with aspect of economics but lie in same cluster (i.e. non economics cluster) after performing economics-oriented clustering.

In an ideal economics-oriented clustering, a non-economics cluster will be produced including all documents which do not associate with economics and each of the other clusters corresponds to a single content of economics.

C. Framework of Aspect-Oriented Document Clustering Approach

The pseudocode and framework of our approach for utilizing Wikipedia in aspect-oriented document clustering is presented in Fig. 2 and 3 respectively.

In this approach, we first, enrich concept of the aspect using Wikipedia based knowledge (line1). Different type of data including redirects, categories, links and the contents of Wikipedia articles are used to find features of the aspect. After that, we scan text documents to extract these features and represent them based on the features in vector space model (line 2). Aspect based document are the input of clustering algorithm for the last step. Finally, documents are clustered with basic k-means, a well-known partitioning clustering algorithm (line 4).

Aspect-oriented document clustering approach using Wikipedia
……
Input**: Document Collection **D**, **Wikipedia**, **Aspect
***Output**: Aspect-Oriented Clustered Documents*
Begin

***1**:Enriched_Aspect←**Enrich_Aspect** (Wikipedia, Aspect)*
2**:**for** each d ϵD **do

***3**:d_Aspect←**Represent_Document_Based_On_Aspect**(d,Enriched_A*

Fig. 2. Pseudocode of aspect-oriented clustering approach.

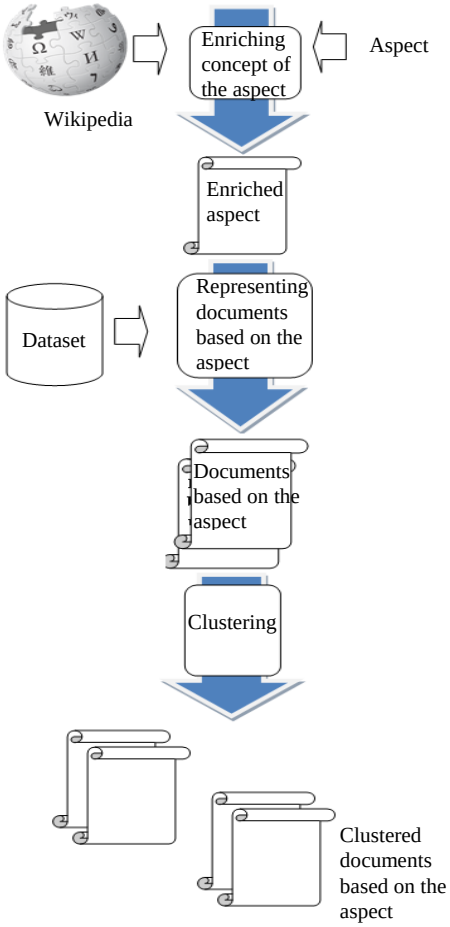

Fig. 3. Framework of aspect-oriented clustering approach.

D. Enriching Concept of the Aspect

The main issue of concern to us is using Wikipedia data to enrich concept of the interested aspect leading to achieve its features. The pseudocode reported in Fig. 4 shows the enrichment process. In this process following data are extracted as aspect`s features:

1) The title, redirects, contents and categories of article of the aspect (lines 1-4)
2) The children of those aspect`s categories which have same title with those articles that have in and also out

link with article of the aspect (lines 5-10).

3) The title, redirects and contents of those articles which have in or out link from/to article of the aspect and their maximum category distance is two (lines 11-14). As Wikipedia has a hierarchical category structure, the distance between a category and its children is equal to one and so on.
4) The children of those aspect`s categories which have the same title with the articles that have in or out link from/to the article of the aspect and their maximum category distance is two (lines 15-18).

```
Enriching concept of the aspect
……
Input: Wikipedia, Aspect
Output: Enriched_Aspect
Begin
1: Enriched_Aspect←Aspect
2: Redirects←Get_Redirects(Aspect)
3: Contents←Get_Contents(Aspect)
4: Cats←Get_Categories(Aspect)
5: Children←Get_Children(Cats)
6: In_And_Out_Links←Get_ In_And_Out_Links(Aspect)
7: for each l ϵ In_And_Out_Links do
8:   for each c ϵ Children do
9:     if (c=l)
10:    Enriched_Aspect←c+ Enriched_Aspect
11: In_Or_Out_Links←Get_ In_Or_Out_Links(Aspect)
12: for each l ϵ In_Or_Out_Links do
13:  if (Get_Category_Dis(l,Aspect)<=2)
14:          Enriched_Aspect←Get_Redirects(l)+Get_Contents(l)+l+
Enriched_Aspect
15: for each l ϵ In_Or_Out_Links do
16:  for each c ϵ Cats do
17:   if(c=l and Get_Category_Dis(l,Aspect)<=2)
18:     Enriched_Aspect←Get_Children(c)+ Enriched_Aspect
19: Enriched_Aspect←Redirects+Contents+Cats+ Enriched_Aspect
```

Fig. 4. Pseudocode of enriching concept of the aspect

E. Representing Documents Based on the Aspect

In this approach, we represent documents based on the interested aspect in vector space model. First, documents are scanned to extract aspect`s features then they will be represented based on four schemas: *concept (maximum distance=1), concept (maximum distance=2), concept-compressed concept (maximum distance=1) and concept-compressed concept (maximum distance=2)*. These schemas are described in table I.

TABLE I: DESCRIPTION OF DOCUMENT REPRESENTATION SCHEMAS

Schema	*Description*
Concept: • Maximum distance= 1 • Maximum distance= 2	Document representation is based on aspect`s features that are extracted from document while category distance between aspect`s article and its linked articles is no more than 2 in enrichment step.
Concept-Compressed Concept: • Maximum distance= 1 • Maximum distance= 2	Document representation is based on aspect`s features which are extracted from document and their compressed version. A compressed phrase is created by ignoring spaces between its terms. For example, *UnitedNations* is the compressed version of United Nations. Also, category distance between aspect`s article and its linked articles is no more than 2 in enrichment step.

F. Clustering

For this step, basic K-Means algorithm [9], a well-known partitioning algorithm is chosen to cluster documents. Documents are represented in vector space model and cosine distance is the similarity measure in the algorithm.

IV. EXPERIMENTS

A. Wikipedia Data

Wikipedia dump files by 16 Sept. 2010, downloadable from http://dumps.wikimedia.org/ are utilized for enriching concept of the aspect.

B. Database Used

In order to evaluate our approach it is needed to cluster documents based on an arbitrary aspect. We choose two issues as interested aspects; each one has adequate number of related documents in a text dataset. The well related categories with the aspect are selected to test from 20 Newsgroups [10] and Reuters 21578 [11] datasets. Actually, categories are subordinate concepts of the aspect.

1) Reuters 21578

For this dataset, we choose *petroleum* as interested aspect for document clustering. Table II Shows the categories related to *petroleum* in Reuters dataset. Each category has at least 10 documents while they are not belonged to more than one category.

TABLE II: SELECTED CATEGORIES FROM REUTERS 21578

Related categories with petroleum aspect in Reuters 21578	
Category	***Number of documents***
Crude Oil	408
Gas Oil	14
Fuel Oil	11
Gasoline	22
Natural Gas	45
Petro-Chemicals	21

2) 20Newsgroups

We select five categories related with *computer* aspect from 20 Newsgroup dataset. Table III lists these categories information.

TABLE III: SELECTED CATEGORIES FROM 20NEWSGROUPS

Related categories with computer aspect in 20Newsgroups	
Category	***Number of documents***
comp. graphics	1000
comp.os.ms-windows.misc	1000
comp.sys.ibm.pc.hardware	1000
comp.sys.mac.hardware	1000
comp.windows.x	1000

C. Tools

For the task of text clustering, we use Dragon toolkit [12], a Java open source application and JWPL [13], Java Wikipedia Library, to utilize Wikipedia xml dumps.

D. Evaluation metrics

In the experiment step, F-Score, Purity and Entropy are used to measure the clustering performance. Entropy exhibits the purity of clusters and is weighted sum of the entropy of all clusters which is defined for cluster i as

$$E_i = -\sum_i P_{i,j} log(P_{i,j}). \quad (1)$$

where $P_{i,j}$ is the probability that a member of cluster i belongs to class j. F-Score is a composite measure of Precision and Recall scores that will be 1 for a perfect clustering and is defined as follows:

$$F-Score=\frac{2\times Precision\times Recall}{Precision+Recall}. \quad (2)$$

Purity is related to the precision measure. It is computed by assigning each cluster to the class which is most frequent in the cluster, and then measuring the accuracy of this assignment [14]. Formally:

$$Purity=\frac{1}{|Documents|}\sum_{clusters} max|Cluster\cap Class|. \quad (3)$$

E. Results

As seen in Table IV cluster Purity, F-Score and Entropy for baseline representation (disregarding the aspect) and the four document representation schemas are represented. A bold font is used to express F-Score and Purity scores higher and Entropy scores lower than baseline. Table V shows improvement of aspect-oriented clustering approach.

Also, Fig. 5 and 6 show a better comparison of the evaluation metrics values for the two tested datasets. It can be observed from Table V and Fig. 6, that the proposed approach has good effect on 20 Newsgroups dataset in all cases in terms of F-Score, Purity and Entropy. But for Reuters 21578 dataset, the approach is effective only for schemas with maximum distance of length two in terms of all three evaluation metrics (see Fig. 5). We see that representing documents based on concept-compressed concept schema, while the concept of aspect has been enriched with in-out links with maximum category distance of length two, changes evaluation metric scores effectively and considerably.

TABLE IV: EVALUATION RESULTS FOR ASPECT-ORIENTED DOCUMENT CLUSTERING

20Newsgroups (computer)			
Document representation schema	***Entropy***	***Purity***	***F-Score***
Baseline	1.244	0.430	0.342
Concept Max distance=1	**1.217**	**0.522**	**0.477**
Concept Max distance=2	**1.133**	**0.514**	**0.491**
Concept-Compressed Concept Max distance=1	1.248	**0.501**	**0.463**
Concept-Compressed Concept Max distance=2	**1.20**	**0.531**	**0.502**
Reuters 21578 (petroleum)			
Document representation schema	***Entropy***	***Purity***	***F-Score***
Baseline	0.639	0.806	0.463
Concept Max distance=1	**0.582**	0.786	0.392
Concept Max distance=2	**0.507**	**0.815**	**0.538**
Concept-Compressed Concept Max distance=1	**0.583**	0.788	0.365
Concept-Compressed Concept Max distance=2	**0.466**	**0.833**	**0.717**

TABLE V: PERFORMANCE OF CLUSTERING RESULTS

20Newsgroups (computer)			
Document representation schema	***Entropy***	***Purity***	***F-Score***
Concept Max distance=1	**-0.027**	**+0.092**	**+0.135**
Concept Max distance=2	**-0.111**	**+0.084**	**+0.149**
Concept-Compressed Concept Max distance=1	+0.004	**+0.071**	**+0.121**
Concept-Compressed Concept Max distance=2	**-0.044**	**+0.101**	**+0.16**
Reuters 21578 (petroleum)			
Document representation schema	***Entropy***	***Purity***	***F-Score***
Concept Max distance=1	**-0.057**	-0.02	-0.071
Concept Max distance=2	**-0.132**	**+0.009**	**+0.075**
Concept-Compressed Concept Max distance=1	**-0.056**	-0.018	-0.098
Concept-Compressed Concept Max distance=2	**-0.173**	**+0.027**	**+0.254**

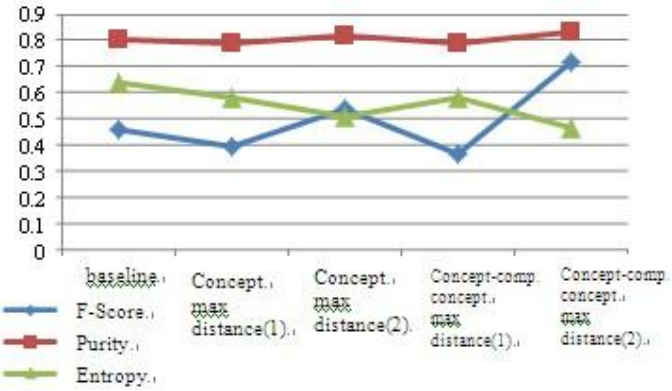

Fig. 5. Diagram of evaluation metrics for petroleum-oriented clustering.

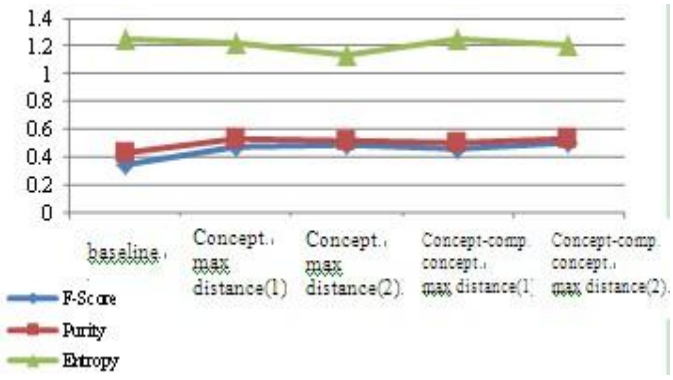

Fig. 6. Diagram of evaluation metrics for computer-oriented clustering.

Table VI shows that dimensionality is reduced by aspect-oriented clustering due to represent documents only based on corresponding aspect`s features.

TABLE VI: DIMENSIONALITY REDUCTION.

Datasets	basic	Concept-Compressed Concept Max distance=2	improvement
20Newsgroup	79305	985	-98.75%
Reuters 21578	10019	320	-96.8%

V. CONCLUSION AND FUTURE WORK

We have introduced a new clustering approach guided by a special point of view denoted aspect in this paper. To enrich concept of the aspect different types of Wikipedia data have been used. Then we have evaluated the proposed approach on two text datasets with three evaluation metrics, F-Score, Purity and Entropy. The Results demonstrate that aspect-oriented clustering enhances clustering performance significantly for those documents which can be equivalent to retrieved documents from aspect based queries. Moreover, clustering speed increases as the dimensionality of text document vectors decreases due to represent them based on only features of the aspect but not all collection terms. Due to the approach`s dependency on the aspect, mining it hidden in user`s queries, improves approach`s applicability. For the queries include more than one aspect, a nested aspect-oriented clustering can be performed. Such an example like ”increasing oil price from the aspect of Middle East developments and its effects from the view point of economics” .

ACKNOWLEDGMENT

Marjan Hosseinia thanks Amir Hossein Jadidinejad for his helpful comments to improve this project.

REFERENCES

[1] R. Feldman and J. Sanger, *The Text Mining Handbook*. Cambridge University Press, 2007.

[2] From Wikipedia, the free encyclopedia. [Online]. Available: http://www.wikipedia.org/

[3] Baboons in Their Natural Habitat (Britannica.com) – YouTube. [Online]. Available: http://www.britannica.com/

[4] O. Medelyan, D. Milne, C. Legg, and I. H. Witten, “Mining meaning from Wikipedia,” *International Journal of Human-Computer Studies*, USA, vol. 67 Issue 9, September 2009.

[5] X. Hu, X. Zhang, C. Lu, E. K. Park, and X. Zhou, “Exploiting Wikipedia as external knowledge for document clustering,” in *Proc. of the 15th ACM SIGKDD*, 2009.

[6] M. Hosseinia, K. Badie, and A. Moeini, “Enhancing performance of document clustering using Wikipedia articles information,” in *Proc. of 2nd International Conference on Contemporary Issues in Computer and Information Sciences (CICIS 2011)*, Zanjan, 2011.

[7] S. Banerjee, K. Ramanathan, and A. Gupta, “Clustering short texts using Wikipedia,” In *Proc. of the 30th Annual International ACM SIGIR Conf. on Research and Development in Information Retrieval*,2007.

[8] N. Kumar, V. V. B. Vemula, and K. Srinathan, “Exploiting N-gram Importance and Wkipedia based additional knowledge for improvements in GAAC based document clustering,” In *Proc. of International Conference on Knowledge Discovery and Information Retrieval*, Valencia, Spain, 2010.

[9] A. K. Jain, “Data clustering: 50 years beyond K-means,” *Pattern Recognition Letters*, vol. 31, June 2010, pp. 651-666.

[10] 20Newsgroups. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups.

[11] Reuters21578. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Reuters21578+Text+Categorization+Collection.

[12] X. Zhou, X. Zhang, and X. Hu, “Dragon toolkit: incorporating auto-learned semantic knowledge into large-scale text retrieval and mining,” In *Proc. of the 19th IEEE International Conf. on Tools with Artificial Intelligence (ICTAI)*, October 2007, Patras, Greece.

[13] T. Zesch, C. Muller, and I. Gurevych, “Extracting lexical semantic knowledge from Wikipedia and Wiktionary,” In *Proc. of the Language Resources and Evaluation Conf.*, Morocco, 2008.

[14] Introduction to Information Retrieval - The Stanford NLP – Stanford. [Online]. Available: http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.htm