A New Sub-topic Clustering Method Based on Semi-supervised Learning

Xiaodan Xu
College of Mathematics, Physics and Information Engineering, Zhejiang Normal University
Jinhua, China
Email:xuxiaodan@zjnu.cn

***Abstract*—Sub-topic clustering is a crucial step in multi-document summarization. The traditional k-means clustering method is not effective for topic clustering because the number of clusters k must be given in advance. This paper describes a new method for sub-topic clustering based on semi-supervised learning: the method firstly partition the set of sentences into disjoint subsets, each of which contained sentences covering exactly one topic, and labels the sentences which have high scores in the topic, then use the method of constrained-k-means to decide the number of topics, and finally get the sub-topic sets by k-Means clustering. This algorithm can dynamically generate the number of k-means clustering, and the experiment result indicates that the accuracy of clustering is improved.**

***Index Terms*—Sub-topic clustering, semantic distance, semi-supervised learning, k-means clustering**

I. INTRODUCTION

With the continuing growth of online information, it has become increasingly important to provide improved mechanisms to find and present textual information effectively. Conventional IR systems find and rank documents based on maximizing relevance to the user query. Some systems also include sub-document relevance assessments and convey this information to the user. More recently, Single document summarization systems provide and automated generic abstract or a query relevant summary. However, large scale IR and summarization have not yet been truly integrated, and the functionality challenges on a summarization system are greater in a true IR or topic-detection context.

Consider the situation where the user issues a search query, for instance on a news topic, and the retrieval system finds hundreds of closely-ranked documents in response. Many of these documents are likely to repeat much the same information, while differing in certain parts. Summaries of the individual documents would help, but are likely to be very similar to each other, unless the summarization system takes into account other summaries that have already been generated. Multi-document summarizations are likely t b essential in such situations. Ideally, multi-document summaries should contain the key shared relevant information among all the documents only once, plus other information unique to some of the individual documents that are directly relevant to the user's query.

Though many of the same techniques used in single-document summarization can also be used in multi-document summarization, there are some significant difference such as the degree of redundancy, the compression ratio, the co-reference problem in summarization presents.

At present, the method based on text clustering is used in multi-document summarization and gets the good results. Reference [1] R.Radev present a multi-document summarizer, called MEAD, which generates summaries using cluster centurions produced by a topic detection and tracking system. MEAD uses information from the cancroids of the clusters to select sentences that are most likely to be relevant to the cluster topic. Reference [2]Endre Boros get the multi-document summaries by utilizing complete sentences from the documents in the collection. In this method, classic clustering techniques were employed in an attempt to partition the set of sentences into disjoint subsets or clusters, each of which contained sentences covering exactly one topic. Clusters are ranked by their similarity with the vector of the term frequencies of all terms appearing in the documents to be summarized. In this method, the similar sentences in multi-document set are combined into one class, each class is one topic of multi-document set, and then multi-document set can be composed of sub-topic sets. Reference [3][4][5] described this method. Therefore the sub-topic detecting is important. Usually, there are two methods for detecting the sub-topics: one is based on hierarchy clustering and the other is based on k-means clustering. The hierarchy clustering method needs an end-clustering threshold, which is hard to decide. The traditional k-means clustering must be given in advance the number of clusters k, but in the actual cases, k is difficult to establish; In addition, traditional k-means algorithm has powerful local search capability, but easily

Manuscript received September 5,2011 ; revised December 19,2011 ; accepted January 1,2012

doi:10.4304/jcp.7.10.2471-2478

falls into local optimum. Genetic algorithm can get the global optimal solution, but the convergence is fast.

In view of this, this paper presents a new method of sub-topic clustering based on semi-supervised learning. The algorithm first partition the set of sentences into disjoint subsets or clusters, each of which contained sentences covering exactly one topic, and labels the sentences which have high scores in the topic, then use the method of constrained-means to decide the number of topics, and finally get the topic sets by k-Means clustering.

The rest of the paper is organized as follows. In section 2 relationship between the approaches here proposed and relevant literature is presented. In section 3 the semantic distance between sentences and the algorithm of sub-topic clustering are formally described. Section 4 reports experimental results and comparison with related methods. Finally, in section 5 conclusions are drawn.

II. RELATED WORK

The literature related to this work can be grouped into two main categories: semi-supervised clustering, the evaluation method of the sub-topic detecting and multi-document summarization.

In many machine learning domains (e.g. text processing, bioinformatics), there is a large supply of unlabeled data but limited labeled data, which can be expensive to generate. Consequently, semi-supervised learning, learning from a combination of both labeled and unlabeled data, has become a topic of significant recent interest.

Clustering is an unsupervised learning problem, which tries to group a set of points into clusters such that points in the same cluster are more similar to each other than point s in different clusters, under a particular similarity metric. Clustering problems can also be categorized as generative or discriminative. In the generative clustering model, parametric form of data generation is assumed, and the goal in the maximum likelihood formulation is to find the parameters that maximize the probability of generation of the data given the model. In the most general formulation, the number of clusters K is also considered to be an unknown parameter.

Semi-supervised clustering, which uses class labels or pair wise constraints on some examples to aid unsupervised clustering, has been the focus of several recent projects. Semi-supervised clustering can group data using the categories in the initial labeled data as well as extend and modify the existing set of categories as needed to reflect other regularities in the data.

Existing methods for semi-supervised clustering fall into two general approaches: search-based and similarity-based methods. In search-based approaches, the clustering algorithm itself is modified so that user-provided labels or constraints are used to bias the search for an appropriate partitioning. In similarity-based approaches, an existing clustering algorithm that uses a similarity metric is employed; however, the

Similarity metric is first trained to satisfy the labels or constraints in the supervised data. Reference[6][7][8][9]Several similarity metrics have been used for similarity-based semi-supervised clustering, including string-edit distance trained using EM,Jensen-Shannon divergence trained using gradient descent, Euclidean distance modified by a shortest-path algorithm,or Mahalanobis distances trained using Convex optimization.Several clustering algorithms using trained similarity metrics have been employed for semi-supervised clustering, including single-link and complete-link and KMeans.

Reference [10][11]Basu proposed two algorithms for semi-supervised clustering with labeled data:seeded KMeans(S-KMeans)and constrained KMeans(C-KMeans). In S-KMeans, the seed clustering is used to initialize the KMeans algorithm. Thus, rather than initializing KMeans from K random means, the centroid of the *h*th cluster is initialized with the centroid of the *h*th partition S_h of the seed set. In C-KMeans, the seed clustering is used to initialize the KMeans algorithm as described for the S-KMeans algorithm. However, in the subsequent step, the cluster memberships of the data points in the seed set are not recomputed in the assign_cluster steps of the algorithm-the cluster labels of the seed data are kept unchanged, and only the labels of the non-seed data are re-estimated. C-KMeans seeds the KMeans algorithm with the user-specified labeled data and keeps that labeling unchanged throughout the algorithm. In S-KMeans, the user-specified labeling of the seed data may be changed in the course of the algorithm. C-KMeans is appropriate when the inintial seed labeling is noise-free, or if the user does not want the labels on the seed data the change, whereas S-KMeans is appropriate in the presence of noisy seeds.

Chinese researchers also get some progress in semi-supervised clustering.

Reference [12] YIN Xuesong presents a discriminative semi-supervised clustering analysis algorithm with pairwise constraints, called DSCA, which effectively utilizes supervised information to integrate dimensionality reduction and clustering. Reference [13] Wang Ling proposed a density-sensitive semi-supervised spectral clustering algorithm (DS-SSC), which incorporate the pairwise constraints knowledge and space consistency prior knowledge into original spectral clustering. Reference [14]Peng Yan proposed a semi-supervised canonical correlation analysis algorithm called Semi-CCA, which used supervision information in the form of pair wise constraints in canonical correlation analysis. Reference[16] Jin Jun described a semi-supervised robust online clustering algorithm called Semi-ROC, which introduced supervision information in the form of class labels into the previously proposed robust online clustering. Reference[17] Wang HJ proposed a semi-supervised cluster ensemble (SCE) model based on both semi-supervised learning and ensemble learning technologies.

In recent years, peoples pay more attention to the standard test sets and large scale evaluations. Two workshops on Automatic Summarization were held: the Document Understanding Conference (DUC) sponsored by the National Institute of Standards and Technology (NIST) started in 2001 in the United States. The Text Summarization Challenge (TSC) task under the NTCIR (NII-NACSIS Test Collection for IR Systems) project started in 2000 in Japan. DUC and TSC both aim to compile standard training and test collections that can be shared among researchers and to provide common and large scale evaluations in single and multiple document summarization for their participants.

III. SUB-TOPIC CLUSTERING

In this paper, D is a document collection, $D=\{d_i|i=1,2,\ldots n\}$, and d_i is a sentence collection: $d_i=\{s_{i,k}|k=1,2,\ldots.m\}$,so multi-document set can be described as the set of sentences $s_{i,l}$: $D = s_{i,l} \mid s_{i,l} \in d_i$,The sentences which have the same meaning are composed to one topic T_i, the multi-document set can also describe as the set of topics: $D=\{T_i|i=1,2\ldots k\}$.by this way, multi-document set is the sets of many sub-topics which describe the articles from different aspect. It is useful to improve the quality of multi-document summarization.

Compared with ordinary text file, the Web page includes a large number of additional information, such as html tags, script, internet link, navigating, copyright. This non-text information may influence the speed and quality of the abstraction and must be filtered before making the abstraction. Because some of the html tags (such as <H1>, Title>, etc) provide useful information for summarization, these useful tags should be kept while cleaning the web pages.

In order to clean the web pages, the following strategy is resented: Firstly, establish a regular expression of text block, and then withdraw the smallest text block which contains the main text information by expression matching. Secondly, combine pattern matching with heuristic rules to clean "the noise" and keep the useful html tags. Finally get the text information.

The so-called automatic summarization method is to detect the sub-topics and deduce the abstract from the different topics by combining some key sentences. It is favorable to understand the logic and fundamental framework of the article and so that the abstract can reflect the contents more correctly and comprehensively. We are going to discuss the sub-topic detecting as the following two steps: calculate the semantic similarity of sentences and detecting the sub-topics based on semi-supervised clustering.

A. Calculate the Semantic Distance between Sentences

Chinese language is different from English in the structure. It does not have the obvious separation symbol between the words, so it need word parsing before further processing. The low-frequency words (only appear once) and some common words such as "the"," and" , "at" are filtered because they contain little information, and the remaining words which called practicable words are used to calculate the semantic distance.

We will use the following formula to calculate the semantic distance between two sentences A and B.

$$D(A,B) = \frac{a}{a + S(A,B)} \qquad (1)$$

Among that, D(A,B) is the semantic distance between sentence A and sentence B, S (A, B) is the semantic similarity of sentences A and B. usually, the higher the similarity between sentences, the shorter their semantic distance.

Accordingly, the following formula is used to calculate the semantic similarity of two sentences:

$$S(A,B) = \frac{\sum_{i=1}^{m} S(a_i, B) + \sum_{j=1}^{n} S(b_j, A)}{m+n} \qquad (2)$$

In the formula (2), sentence A contains words a_1, $a_2\ldots\ldots a_m$, and sentence B contains the words:$b_1,b_2\ldots\ldots b_n$. S (a_i， B) is to calculate the semantic similarity between the word a_i and the sentence B, the formula for it would be :$S(a_i, B)=\text{Max}(S(a_i,b_1),S(a_i, b_2),\ldots..S(a_i, b_n))$.the same for S($b_j$， A).

Obviously, $S(a_i,b_j)$ is to calculate the semantic similarity of word between a_i and b_j.

$$S(a_i, b_j) = \frac{\alpha}{\alpha + D(a_i, b_j)} \qquad (3)$$

In the formula, $D(a_i,b_j)$ is the semantic distance between a_i and b_j. We use a thesaurus dictionary to calculate the value of $S(a_i,b_j)$, and the dictionary is provided by Research Center for Information Retrieval (HIT-CIR) of Hrbin Institute of Technology. The structure of thesaurus dictionary and its levels are as follows:

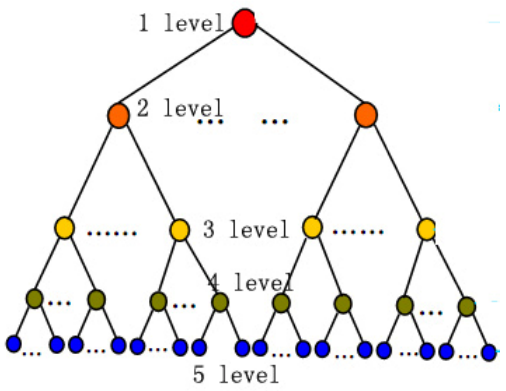

Figure 1. Tree structure of thesaurus dictionary

For the words in dictionary, there are five levels of semantics, which use different mark. The first level use capital letters (A-Z), the second level use small letters(a-z), the third level use numbers(00-99), the fourth and fifth level additional can provide more information. Each word is encoded and arranged according to their semantic relations, for example, the encode of "peach" is Bh07A28.

We use the following formula to calculate the value of $D(a_i,b_j)$:

$$D(a_i, b_j)=2*(6-n) \quad (4)$$
$$(2<=n<=6)$$

As for each word has a semantic code, n is the start position which the semantic code is different between two words. For example, in the dictionary, the code for "peach" is Bh07A28, and that for "watermelon" is BhO7A56, so the value of n is 5 and the semantic distance between the two words can be calculated like D(peach,watermelon)=2*(6-5)=2. Exceptional, for the situation n=1,if the two words belong to the noun class or verb class, their semantic distance would be $D(a_i, b_j)=12$,else $D(a_i, b_j)= +\infty$.

B. Cluster the sub- topics

As the semantic distance of each sentence is calculated, the sentences which have small distance will be clustered into one class. Each class would be a sub topic.

Hierarchical clustering was used for finding the initial clusters:

1. Start with each sentence being a cluster of size 1
2. Cal Calculate the distance between each cluster and sort a list of this

Information so the "closest" clusters are at the top .

3. Pick the two clusters which are "closest" and merge them into a new cluster.
4. Delete the two "closest" clusters and any references to them in the distance list.
5. Go to 2.
6. Stop when have trimmed down to m clusters.

Non-hierarchical clustering, specifically k-means clustering method is given the m clusters as a starting point, with a target of trimming the number of clusters to n(n<m).Since k-means may terminate with more than the target of 10 clusters, the 10 clusters with the most sentence in them are utilized.

From the above we can see that the traditional approach based on hierarchical clustering need an end-clustering threshold, which is hard to decide; the k-means clustering need the Initial Value of k, while the number of sub topics is unpredictable. The method of semi-supervised clustering we will describe in the following can effectively overcome these shortcomings and obtain good results.

1) K-means clustering based on semi-supervised

The basic idea of semi-supervised learning is to use the labeled data to predict the unlabeled data. In semi-supervised clustering; some label level or instance level supervised information is used along with the unlabeled data in order to obtain a better clustering result.

Many of the existing semi-supervised clustering algorithms are based on the traditional clustering algorithm .On behalf of the algorithm is the semi-supervised k-means algorithm which developed from the classical k-means algorithm.

Given a set of observations ($\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$), where each observation is a *d*-dimensional real vector, *k*-means clustering aims to partition the *n* observations into *k* sets ($k \le n$) $\mathbf{C} = \{C_1, C_2, \ldots, C_k\}$ so as to minimize the within-cluster sum of squares.

TABLE I.
WORDS CODE LIST OF THE DICTIONARY

level	Symbols used	For example	sementic
1	Capital letters	B	Main category
2	Small letters	h	Sub category
3	Double digit	07	Detail category
4	Capital letters	A	Words cluster
5	Double digit	28	Detail words cluster

The average value of class k can be described as $m_k = \frac{1}{N_k}\sum_{i=1}^{N_K} x_i$, k=1,…K, and the Objective function of K-means clustering based on Euclidean distance[18] is as followings:

$$J = \sum_{k=1}^{K}\sum_{i=1}^{N_K} \| x_i - m_k \|^2 \quad (5)$$

K-means clustering acquires the number of clusters in advance. The random selection of the initial cluster centers will result in the instability and K-means clustering algorithm will be terminated in access to a local optimum value.

K-means clustering method usually takes k as the initial value on the condition that the number of clusters is given in advance. The following two graphs show the different conditions of data distribution with two classes.

The coordinate system in fig2, fig3 and fig4 is based on the distance between two data. The fig3 shows the data sets are disjointed and fig4 shows the data set are intersected.

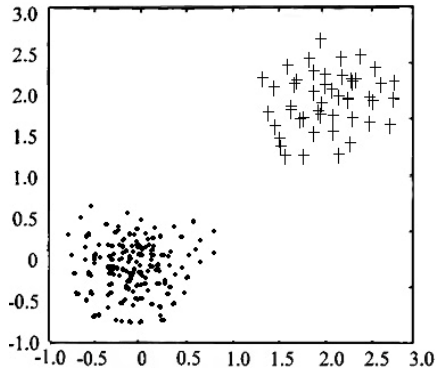

Figure 2. Distributing graph of two disjoint data sets

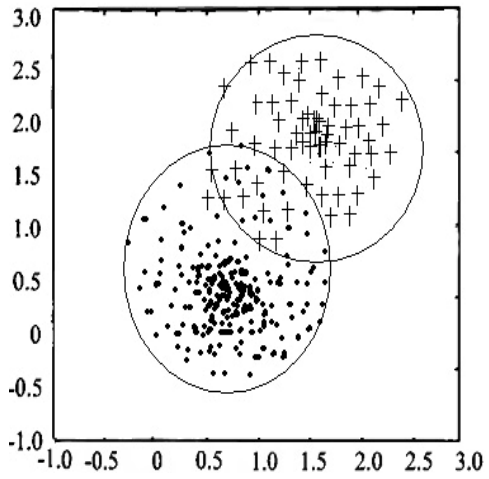

Figure 3. Distributing graph of two intersection data sets

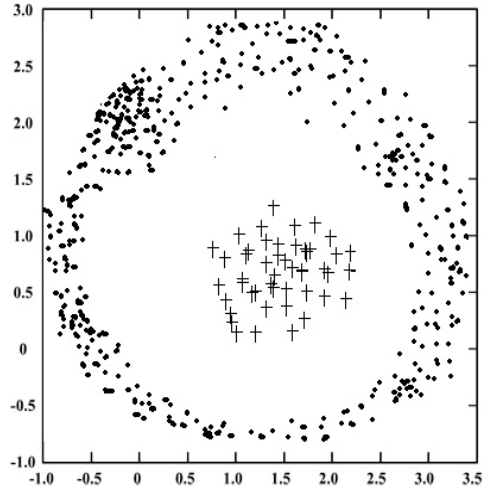

Figure 4. Distributing graph of data sets with non-covex shapes

Figure2 shows that the k-means algorithm can get good effort for disjoint data sets when k=2, while in figure3 and figure4 we can see that the effect of clustering will drop down. Data shown in fig3 contains two categories, if k=2, data at the cross position will be assigned to error categories, and if we take the value of k greater than 2, more data can meet the rules that data inside the class have a strong similarity and between class have low similarity.

However, the increased K value method requires the following two judgments: one is what value k takes can maximize accuracy of clustering on the existing foundation; the second is how to judge the redundant marked categories when the clustering is completed.

In order to solve the problem, we try to use the semi-supervised method to determine the initial value of k:

Suppose a complete data set contains a few labeled data set L (monitoring information) and unlabeled data set U, i and j is the mark of labeled set. We first use constrained-K-means clustering method for the data set {L,U} on condition that k=2, then calculate the number(N) of incorrectly labeled data in L when different values of k, the k that make N to obtain the minimum value should be the best initial value. The formula is as follows:

$$N = \sum_{c=1}^{K} \min(n_{ic}, n_{jc}) \quad (6)$$

In (6), c is the class number(c=1,2…K), n_{ic} is the count of data that be marked as class i , n_{jc} is the count of data that be marked as class j. The maximum value of k is adopted as 6 according to the experience.

The constrained KMeans method can be described as follows[20]:

Input: set of data points $X=\{x_1,\ldots x_n\}$,number of clusters K, set $\mathbf{S}= \{S_1, S_2, \ldots, S_k\}$ of initial seeds

Output: disjoint k partitioning $\{x_h\}$(h=1..,k) of X such that the KMeans objective function is optimized

1.Initialize clusters: $u_h^{(0)} \leftarrow \frac{1}{|s_h|} \sum_{x \in s_h} x$, for h=1,…K; $t \leftarrow 0$

2.repeat until convergence

2a. assign_cluster: for x∈S,if x∈Sh assign x to the cluster h(ie.,set $X_h^{(t+1)}$).For $x \notin S$,assign x to the cluster h*(ie.,set $X_{h^*}^{(t+1)}$),for $h^* = \arg\min_h \left\| x - u_h^{(t)} \right\|^2$

2b. estimate_means: $u_h^{(t+1)} \leftarrow \frac{1}{\left| X_h^{(t+1)} \right|} \sum_{x \in X_h^{(t+1)}} x$

2c. $t \leftarrow (t+1)$

The detail method of sub-topic clustering will be described in the following section 2).

2) the algorithm of semi-supervised clustering

While detecting the sub topics, we use the following method:

(1) Given the number of sub-topics (n=3), get the primal sets of sub-topics by hierarchy clustering, these

sets can be described as T = (T_i) (1 <= i <= n), where T_i is collection of sentences, $T_i = (s_{i,k} \mid k = 1,2, ... m)$;

(2) For each T_i, label the sentences that have high similarity and add them into the labeled data set, then get two sets: labeled data set L which contains few labeled sentences, and unlabeled data set U;

(3) Use constrained-K-means clustering method for the data set {L,U} on condition that k=2, then calculate the number(N) of incorrectly labeled data in L when different values of k, take the k that make N to obtain the minimum value.

(4) Cluster all sentences with k-means given k, and finally get the sub topics.

TABLE II.
SAMPLES IN EXPEREMENT

Theme	Collection number	The average number of sentences
military	2#,3#,12#,14#,15#	320
art	1#,4#,5#	280
amusement	6#,10#,13#	300
education	7#,8#,9#,11#	210

To determine which sentences should be selected to be included in the summary and the order in which they should appear, clusters were ranked by their similarity to the collection term frequency vector. The sentences within each cluster were then ranked by their similarity to their cluster center. And the Sentences in front of the team row of each sub-topic is detected to be included in the summary in turns .

IV. EXPERIMENTS AND EVALUATION

The presented experiment sample comes from people's network in 2001 which includes about 6000 web pages. The data consists of eight classes, namely military, lift, amusement and so on. In the experiment, we extract 15 collections of web pages; each collection contains 5-10 pages. The web pages in one collection refer to the same topic.

The collection of sample in the experiment is list in table Ⅱ.

While detecting the sub-topics, we compare the semi-supervised clustering with the traditional hierarchy clustering, the result shown in tableⅢ.

In this paper, we first get the sub topics by expert, and then take the clustering accuracy P to evaluate the results. Above that P is describes as $P = \frac{N_{righ}}{N_{all}}$., N_{right} refers to the number of sentence that to be correctly classified by our system.

The experimental results show that some documents based on hierarchical classification do not play an effect, there are two reasons: one is that the number of sub-topics pre-given is not the best; the second is that the method based on hierarchical classification can not be backtracking. While the semi-supervised learning method proposed in this paper can be more precise in determine the number of sub-topics, which effectively improve the classification results.

TABLE II.
THE ACCURACY OF CLUSTERING BY TWO METHODS

web pages set	Sub topics number	Hierarchy clustering	Semi-supervised clustering
1#	4	70%	75%
2#	2	80%	85%
3#	3	78%	78%
4#	5	55%	58%
5#	3	66%	70%
6#	5	60%	62%
7#	3	73%	77%
8#	4	70%	78%
9#	5	74%	80%
10#	4	61%	68%
11#	3	81%	88%
12#	3	80%	80%
13#	4	65%	68%
14#	4	71%	80%
15#	3	75%	82%

After sub-topic clustering, the key sentences from different topics are extracted to combine the summary. In order to verify the validity of our approach, we extract 5 web pages sets from the military theme and get the summarization with the following 3 methods:

1. TOP-N method: to get the summary by selecting the former N sentences of each article.

2:MEAD method: it firstly select the key words as the mead of documents, and then measure the importance of one sentence by calculating the similarity between the sentence and the mead, finally get the key sentences as a summary.

3.STSB method: summary based on sub-topic clustering which is provided in this paper. The sentences within each cluster were then ranked by their similarity to their cluster center. one sentence at each iteration was selected to be included in the summary. We use the

precision of summary p to evaluate the effect, p=(the number of correct sentences)/(number of all sentences in summary). We also use the experts' summary as the evaluation standard.

The following paragraph 5 shows the precision of summary with three methods in different compression ratio. The compression ratio are 10%,20%,30%.

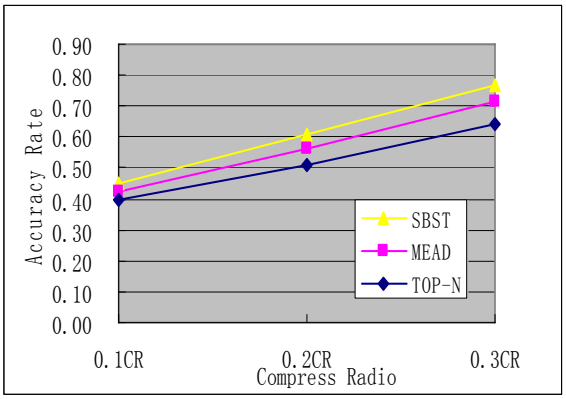

Figure 5. The pricision of summay in different compression ratio

As is shown in figure 5, the quality of most summary is satisfied. However, there is still some redundant information in the summary. In the test, we found that some summary include the sentences with the same meaning, that is because the same important message will be described many times in the sub-topic.

In order to reduce the redundancy, we take the following method: when a sentence is selected into the summary, it must abide by the rules: the key words' coverage rate of summary will be the maximum in all situations, when this sentence is added into the summary.

$$s_{Ti} = \arg\max_{Sij} \frac{|\, Sumword \cup \{Senword_{ij}\} \cap Muldocword \,|}{|\, Muldocword \,|}$$

in the formula ,S_{Ti} is the candidate sentence in sub-topic Ti, *Sumword* is the key words set in summary , and *Muldocword* is the key words set in multi-documents.$Senwords_{ij}$ is the words set in sentence Sij. The purpose is to make the summary cover the content as much as possible. Aided with this strategy, the precision of summary is increased to 80% in figure 6 (SBST').

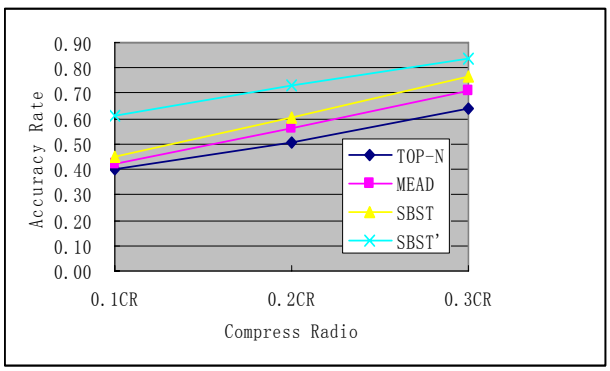

Figure 6. The new pricision of summay in different compression ratio

V. CONCLUSIONS

This paper presents a new method for sub-topic clustering based on semi-supervised clustering. In this method, the semi-supervised clustering is more effective than hierarchical clustering and k-means clustering because it can get the best value of k according to the characteristics of the data. Experiments show that the method is useful. In the next work we will study how to use the technology of natural languages understanding to improve the quality of summary, especially to improve the readability of the summary.

REFERENCES

[1] R. Radev, Hongyan Jing and Malgorzata Budzikowska.2000.Centroid-based summarizaiton of multiple documents:sentence extraction,utility-based evaluation, and user studies.//ANLP/NAACL 2000 Workshop C,2000:21-29

[2] Boros E et al. A clustering based approach to creating multi-document summaries// Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New Orleans, LA , 2001: 34-42

[3] Naomi Daniel,Dragomir Radev,and Timothy Allison.Sub-event based multi-document summarization[A].In:HLTNAACL Workshop on Text Summarization.Edmonton Alberta.Canada.2003:9-16

[4] Pascal Fung,Grace Ngai Combining Optimal Clustering and Hidden Markov Model for Extrative.In:Proceeding s of the ACL 2003 workshop on multilingual summarization and question answering.2003:21-28

[5] Qin Bing,Liu Ting,ChengShanlin.etc.Sentence Optimum Selection for Multi-Document Summarization(in Chinese).Journal of Computer Research and Development.2006.43:1129-1134

[6] Bilenko,M.,&Mooney,R.J..Adaptive duplicate detection using learnable string similarity measures .InProceedings of the Ninth ACMSIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2003),pp.39–48Washington,DC.

[7] Klein,D.,Kamvar,S.D.,&Manning,C..From instance-level constraints to space-level constraints:Making the most of prior knowledge in data clustering.In Proceedings of the The Nineteenth International Conference on Machine Learning(ICML-2002) Sydney,Australia .

[8] Hillel,A.B.,Hertz,T.,Shental,N.,&Weinshall,d. Learning distance functions using equivalence relations. In Proceedings of 20th International Confrerence on Machine Learning(ICML-2003)

[9] Xing,E.P.,Ng,A.Y.,Jordan,M.I.,&Russell,S..Distance metric learning,with application to clustering with side-information.In Advances in Neural Information Processing Systems 15 .MIT Press. 2003.

[10] Basu S,Baneoee A, Moonev R J. Semi, Supervised clustering by seeding ‖ Proc of the 19th International Conference on Machine Learning.Sydney,Australia,2002:19-26.

[11] Basu S,Banerjee A,Mooney R J.Semi-Supervised Clustering by Seeding//Proc of the 19th International conference on Machine Learing.Sydney,Australia,2002:19-26

[12] Yin Xuesong,Hu Enliang,Chen Songcan. Discriminative Semi-Supervised Clustering Anaysis with Pairwise constraints. Journal of Software,2008,19(11):2791-2802(in Chinese).

[13] Wang Ling,Bo Liefeng,Jiao Licheng. Density-Sesitive Semi-Surpervised Spectral Clustering .Journal of Software,2007,18(10):2412-2422(in Chinese)

[14] Xiao Yu,Yu Jian. Semi-Supervised Clustering Based on Affinity Propagation Algorithm. Journey of Software,2008,19(11):2803-2813(in Chinese)

[15] Peng Yan, Zhang Daoqiang. Semi-Supervised Canonial Analysis Alogorithm. Journey of Software,2008,19(11):2822-2832(in Chinese)

[16] Jin Jun, Zhang Daoqiang. Semi-Supervised Robust On-line Clustering Alogorithm .Journal of Computer Research and Development,2008:496-502(in Chinese)

[17] Wang Hongjun,Li Zhishu,Qi Jianhuai,etc.Semi-Supervised Cluster Ensemble Model Based on Bayesian Network. Journey of Software,2010,21(11):2814-2824(in Chinese)

[18] Li Kunlun.Cao Zheng.etc. Some Developments on Semi-supervised Clustering(in Chinese).PR&AI. 2009.10:735-742

[19] ZHU X J.Semi-supervised Learning Literature Survey[R].Madison:University of Wisconsin,2008

[20] Sugato Basu, Semi-supervised Clustering:learning with limited user feedback, Austin.2003.11

Xiaodan Xu,Jinha,China,1978.Received M.A degree of Software Engineering in National University of Defense Technology in 2005,Changsha,China.
Her reaserch interests include data mining and knowledeg discover,computational intelligence ,and automatic summarization.She has worked in the Department of Mathematic Physics and Imformation Engineering,Zhejiang Normal University.China.

A New Method of Sub Topic Clustering in Multi-Document Summarization,Xu.