# Improving Document Clustering by Utilizing Meta-Data*

**Kam-Fai Wong**
Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong
kfwong@se.cuhk.edu.hk

**Nam-Kiu Chan**
Centre for Innovation and Technology, The Chinese University of Hong Kong
jussie@cintec.cuhk.edu.hk

**Kam-Lai Wong**
Centre for Innovation and Technology, The Chinese University of Hong Kong
klwong@cintec.cuhk.edu.hk

## Abstract

In this paper, we examine how to improve the precision and recall of document clustering by utilizing meta-data. We use meta-data through NewsML tags to assist clustering and show that this approach is effective through experiments on sample news data. Experimental result shows that clustering using NewsML could improve average recall and precision over the same without using NewsML by about 10%. Our algorithm facilitates effective e-business for the news media and publishing industry to empower e-business.

## 1 Introduction

Nowadays, people have great demand on knowledge and information, while information overload becoming one serious problem. News media and publishing industry therefore try to suit customers' need by using electronic information management system. Document clustering algorithm has been introduced to group similar documents together for easier searching and reading.

Document clustering algorithm has been widely used in news media and publishing industry, which ensured it effectiveness over manual clustering. With labor cost reduced and time saved, document clustering algorithm provides convenient clustered-news for users.

To improve the accuracy of document clustering algorithm, we suggest to provide more flexible information for each document. Under the hypothesis that document clustering algorithm can get better result with more information about data used, we suggest that using additional meta-data contained in NewsML standard could enhance the performance of document clustering algorithms.

We evaluated the effectiveness of using meta-data in the proposed clustering algorithm. We used Chinese electronic news sources for the evaluation. The experiment showed that using the meta-data provided by NewsML achieved better document clustering.

The remaining of the paper is organized as follows: In Section 2, we give an overview of current document clustering approaches. Section 3, analyses existing problems in document clustering and suggests a solution using NewsML. We show the tags that could be used in the algorithm and how they are handled. The performance of algorithm through experiments, and we will present the experimental results regarding measures in precision and recall. In Section 5, we include a brief summary and discussion on future work.

## 2 Current Approaches

Different document clustering methods have been examined. These conventional clustering methods mainly consist of two parts: construction of a similarity matrix between documents and formulation of clustering algorithm to generate clusters.

### 2.1 Similarity Matrix

The first step of conventional clustering method is to construct a similarity matrix between these documents so as to understand how documents are similar to one another. The constructed similarity matrix will later be used by the clustering algorithm for generating clusters.

* Corresponding Author: Kam-Lai Wong (klwong@cintec.cuhk.edu.hk)

Named entity method (Volk and Clematide, 2001) is one of the widely used approaches for constructing a similarity matrix. Named Entities form the major components in a document. When fundamental entities like Person, Company and Geographical names are detected, the algorithm could understand the content of a document to a certain extent.

Named entity method has also been investigated together with keywords to perform clustering (Lam, Meng, Wong and Yen, 2001). The similarity score is calculated based on the named entity vectors and the keyword vector with weighting parameters to control the degree of emphasis on the corresponding vectors.

Concept terms method (Wong, Lam and Yen, 1999) has been proposed in order to deal with the problem of vocabulary switching. The potentially concept terms are basically the keywords derived from a separated concept generation corpus. Concept terms are selected based on the co-occurrence between a query and a document.

However, named entities approach and concept terms approach contain some limitations: The accuracy of the clustering algorithm would be directly proportional to the accuracy of algorithm. Thus, any error from identifications of named entities or concept terms will adversely affect the quality of the clustering algorithm as well.

N-gram Algorithm (Lee, Cho and Park, 1999) has been introduced in order to avoid the aforementioned limitations. An N-gram is a character sequence of length N extracted from a document. The main idea of the N-gram approach is that the character structure of a term can be used to find semantically similar terms. The approach assumes no prior linguistic knowledge about the text being processed. Moreover, there is no language-specific information used in the N-grams approach, which qualifies this method as a language-independent approach. By using N-grams, frequently appeared terms of each document can be extracted and compared to make the similarity measure.

## 2.2 Clustering Algorithm

Probabilistic method is one of the commonly used methods in document clustering. The aim of probabilistic method is to minimize the heterogeneity in each group with respect to the group representative based on statistical approaches (Estivill-Castro and Yang, 2001). Neural network is also used to perform a cyclic learning process for clustering (Grothkopf, Andernach and Stevens-Rayburn, 1998).

Hierarchical clustering methods include group-average clustering algorithm and single-link clustering algorithm (Johnson and Kargupta, 2000; Tombros, Villa and Van Rijsbergen, 2002). Group average clustering is based on creating a hierarchical tree by initially creating a singleton cluster for each document. The clusters are merged to the parent node until the algorithm goal is achieved. The algorithm merges document pairs in the resulting clusters by merging clusters in a greedy, bottom-up fashion. A divide-and-conquer strategy can be used to balance the cluster quality and computational efficiency.

The basic steps of group-average clustering algorithm are like this: On each iteration, it first divides the current pool of clusters into evenly sized buckets. Group-average clustering is then applied to each bucket locally, merging smaller clusters into larger ones. The time complexity for the algorithm is O(kn), where k is bucket size and n is the number of documents.

Single link clustering (Dunlop, 2000), on the other hand, is based on creating a hierarchical tree by continually inserting an additional node that satisfies the following criteria:

- The new node is currently outside the hierarchy
- Of all similarities between nodes inside and outside the hierarchy, the new node which has the strongest similarity is selected. It is then added to the hierarchy at a level based on how strong the similarity is.

The approach is fairly fast and result in hierarchies where the closest nearest neighbours are at lower levels of the hierarchy. However, it leads to non-balanced clusters, and many node-node comparisons can have the same strength of similarity thus many documents can be linked at the same level in the hierarchy.

The accuracy of the above conventional clustering method, however, is generally low. Therefore a new approach is proposed.


**Figure 1: A sample document in NewsML format containing keywords and subject labels**

## 3 Our Proposed Method

In our proposal, we suggest the following approaches for generating clusters.

1. Use Bi-gram to extract terms from documents
2. Use <KeywordLine> Tag to look up keyword terms for documents
3. Compare terms betweens documents to construct similarity matrix
4. Use <SubjectCode> Tag to group documents to different subjects
5. Adjust similarity matrix by data provided by step 4
6. Apply group-average clustering algorithm to generate clusters

We have chosen to use Bi-gram algorithm to extract terms from documents. The idea of bi-gram Algorithm is similar to N-grams'. The reason of using bi-gram instead of N-gram is that our experiment mainly deals with Chinese (Big5) news. Since Chinese terms are typically formed by two Chinese characters, bi-gram approach is sufficient for this application. Using bi-gram, moreover, would be a more effective approach when handling other two-byte code like Japanese and Korean languages (Lee, Cho and Park, 1999).

Using bi-grams instead of N-grams can reduce system resources. Since hundreds of documents are handled each time, using N-gram would be impractical. For instance, using N-grams approach for a document with M Chinese characters would extract (M-1) + (M-2)+…+[M- (N-1)] terms. This would require more time for comparison than only M-1 terms for bi-grams.

We suggest using NewsML (http://www.newsml.org) in our project. NewsML, which is released by International Press and Telecommunications Council (IPTC) (www.iptc.org ), is an XML-based data format for news that is intended to use for the creation, transfer and delivery of news. All news is created based on NewsML data type definition (DTD) file. NewsML has been widely recognized globally. Noticeable users include Reuters, AFP, and Kyodonews.

A sample document in NewsML format is shown in Figure 1. By using NewsML, a helpful tag called <KeywordLine> can be used in order to look up existing keywords from a document. The keywords, which could highly reflect the main concept of the document, would usually be given by the author of the document and stored in the <KeywordLine> tag under NewsML. Thus, the <KeywordLine> Tag is a useful indicator for keyword extraction.

We used NewsML tagged keywords for comparison. A score is computed for each keyword based on term frequency to reflect the level of importance to the document. A given keyword, however, may exist in more than two Chinese characters. But in order to make easy and accurate comparison, we extracted the keywords in two Chinese characters each and compare them with the terms extracted by the bi-gram algorithm. These keywords would be more representative

than those extracted by bi-gram. Thus, they would be very useful for identifying the news content. We therefore applied higher weighting to these terms. In view of this, we assign the weighting of NewsML tagged keywords ten times more than those extracted using bi-grams.

Besides using only the keywords and terms from a news document, subject of the news is also crucial. For example, news referring to entertainment should not be in the same cluster with one talking about business news. Understanding news semantically to a certain extent can help improve clustering accuracy. <SubjectCode> Tag in NewsML enabled the clustering algorithm to distinguish which subject a particular news is referring to. Although the <SubjectCode> only gives a rough concept of the news, it can significantly improve accuracy.

In our experiment, we ensured that news with different <SubjectCode> would not be put in the same cluster. The similarity measure of two pieces of news with different <SubjectCode> would, therefore, be zero.

## 4 Experimental Results

We assessed the effectiveness of using meta-data for document clustering empirically. We used NewsML meta-data and Hong Kong Chinese news articles for this purpose.

Performance metrics are based on Recall and Precision, which are defined as follows:

| | In event | Not in event |
|---|---|---|
| In cluster | A | B |
| Not in cluster | C | D |

Recall = A / (A+C) if A+C > 0
Precision = A / (A+B) if A+B > 0

To study the effect of NewsML, we calculated the percentage of recall and precision under different threshold values. The threshold value (0-1) is a user-defined variable for clustering. The threshold is the value where documents within a cluster should have a similarity greater than it, and similarity between two documents is calculated by Jaccard's coefficient:

$$\frac{w_a}{w_a + w_b + w_c}$$

where

a is the set of terms present in both documents

b is the set of terms present in document 1 but absent in document 2

c is the set of terms present in document 2 but absent in document 1

$w_a, w_b, w_c$ are the sum of weights of the related set.

A higher threshold represent higher similarity is needed for documents to be put in the same cluster.

Prior to the experiment, we matched each piece of news with specific Subject Codes defined by IPTC (http://www.iptc.org/site/subject-codes/index.html), as well as adding keywords with respect to their contents. After the preparation process, we started our experiment and find out the recall and precision measures as above stated

The main difference between the algorithms with and without NewsML is on the use of the KeywordLine and SubjectCode tags. The evaluation steps are as follows:

1. Non-overlapping clusters are generated by our clustering system
2. Each generated cluster is matched to the most similar sample event using a one-to-one matching method.
3. If the number of clusters generated is less than the number of events in the data set, discard the excessive sample events after matching.
4. If the number of clusters generated is greater than the number of events in the data set, empty events are added to the sample and they are matched with the generated clusters.
5. Calculate recall and precision for each cluster pairs
6. Use the calculated recall and precision, to obtain the macro-average (Yang, Pierce and Carbonell, 1998).
7. Repeat the above steps on different threshold values.

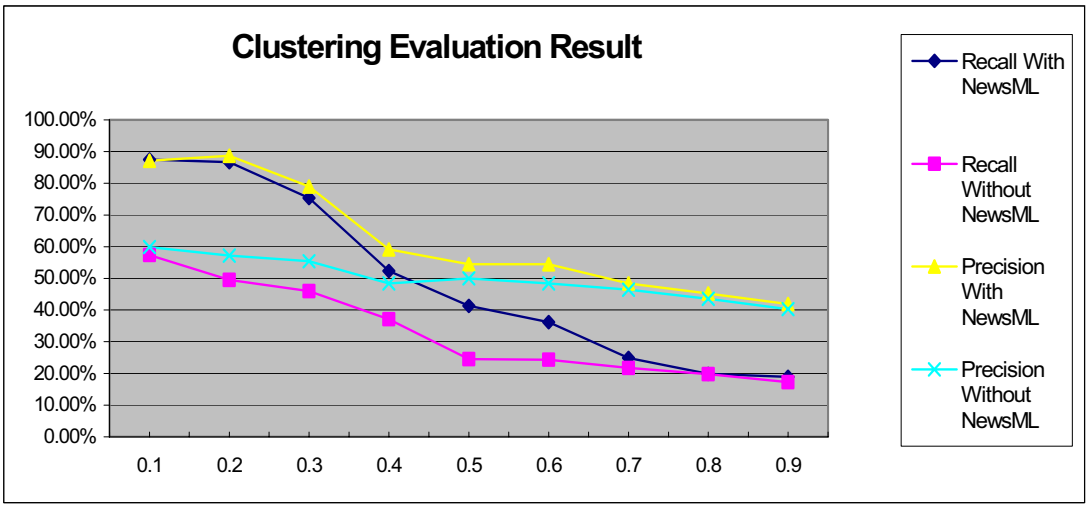


**Figure 2: Clustering evaluation result for 158 news articles from three different publications**

### 4.1 Experiment A: News Articles from Multiple Publications

We began our evaluation with a set of data from three different publications. We used 158 news articles containing 31 distinct events (which was equivalent to 31 non-overlapping clusters) as our training set. The events belonged to different sector including local news, international news, sports news, entertainment news and business news. Based on the IPTC Subject Reference System (www.iptc.org), these news articles were classified into 7 different subjects.

The result is shown in Figure 2 where x-axis represents the threshold and y-axis represents the percentage of the measure.

The graph suggests that both recall and precision values decrease with increasing threshold. The performance of using NewsML is better when the threshold is smaller. The greatest difference occurs when threshold value = 0.2

Moreover, we observe that the variation of the performance is greater when using NewsML. The precision and recall can drop about 40% by changing the threshold. In the contrast, without using NewsML, the result is more steady.

On average, recall and precision NewsML-based algorithm gives 16.15% and 12.09% respectively improvement of its non-NewsML counterpart.

### 4.2 Experiment B: News Articles from the Same Publication

We then carried out another experiment using 503 news articles from one single publication. We tried to classify these news articles manually to events of related news, and then carried out the same approach as the first experiment. These 503 news articles contained 126 distinct events and classified to 12 different subjects with reference to the IPTC Subject Reference System.

We find out how the system performs when dealing with one single publication. As we predict, news from only one single publication often uses same keywords or terms to express the same concept. This simplified clustering.

The result is shown in Figure 3 where x-axis represents the threshold and y-axis represents the percentage of the measure.

The performance of the system decrease when the threshold value increases, which is consistent with the first experiment. From the graph, we observe that the performance is best when the threshold is between 0.2 and 0.4, which means those values could best match human expectation.

The graph also suggests that both recall and precision gives a better result towards using meta-data. On average 8.02% (recall) and 8.92% (precision) improvement over the approach without using NewsML meta-data are achieved.

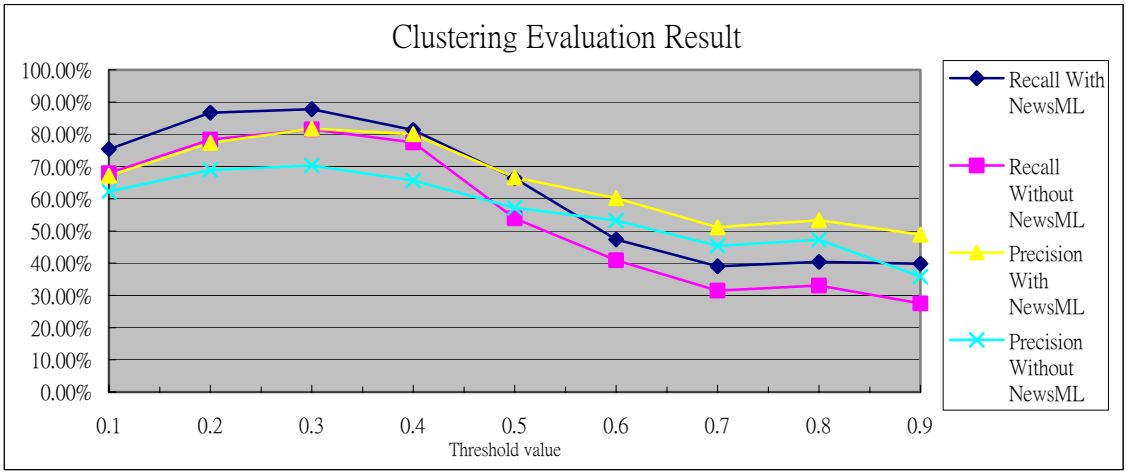


**Figure 3: Clustering evaluation result for 503 news articles from the same publication**

From the results of the two experiments, we have found that using NewsML could improve both the recall and precision of the document clustering algorithm by about 10%, over those without using NewsML.

## 5 Conclusions

In this paper, we demonstrate that the effectiveness of document clustering algorithm could be improved by utilizing meta-data in additional to the original data content. In our experiment, we chose NewsML as a representation of news content with added meta-data. We proposed to use the <KeywordLine> tags and <SubjectCode> tags in NewsML for clustering.

Our proposed document clustering algorithm is a refinement of conventional group-average clustering algorithm and bi-gram algorithm. It has been shown that NewsML could help conventional clustering methods improve both the recall and precision by about 10% on average.

In order to demonstrate the practicality of NewsML for e-business, we have deployed NewsML in developing an application called NewsFocus[1]. News Focus consists of a clustering function. The function is mainly for clustering similar news from three different news sources in Hong Kong. News articles under NewsFocus are clustered by news events.

In the future, we try to use different methods to further improve the clustering performance. Inverted document frequency and cosine-angle formula, for example, have been widely used in terms score calculation and matrix similarity calculation. Top relevance terms can be used as keywords in case of any insufficiency of metadata. We will also try to use other tags in NewsML like <Headline> in supplement with <KeywordLine> and <SubjectCode> to give more metadata information for the document. Weighting parameters may also be applied to show the degree of emphasis on using those metadata.

## References

M. D. Dunlop, *Development and evaluation of clusteing techniques for finding people,* Proceedings of the Third International Conference Basel, Volume 34, 2000

V. Estivill-Castro and J. Yang, *Non-crisp Clustering by Fast, Convergent, and Robust Algorithms*, Principles of Data Mining and Knowledge Discovery, Volume 2168, 2001, pp. 103-114

U. Grothkopf, H. Andernach, S. Stevens-Rayburn, and M. Gomez, *Comparison of Two "Document Similarity Search Engines"*, Library and Information Services in Astronomy III, ASP Conference Series, Volume 153, 1998, pp. 85-92

E. L. Johnson and H. Kargupta, *Collective, Hierarchical Clustering from Distributed, Heterogeneous Data*, Large-Scale Parallel Data Mining, Lecture Notes in Artificial Intelligence, Volume 1759, 2000, pp. 221-244

Lam, W., Meng, H., Wong, K.L. and Yen, J., *Using Contextual Analysis for News Event Detection*, International Journal of Intelligent Systems, 16(4), 2001, pp.525-546

[1] Please visit http://www.cnewsml.org/clustering/jsp/index2.html for demonstration

J. H. Lee, H. Y. Cho and H. R. Park, *N-gram-based indexing for Korean text retrieval*, Information Processing and Management, Volume 35 Number 4, 1999, pp. 427-441

A. Tombros, R. Villa, C. J. Van Rijsbergen, T*he effectiveness of query-specific hierarchic clustering in information retrieval*, Information processing & management, Volume 38, 2002 , pp. 559-582

M. Volk and S. Clematide: *Learn-Filter-Apply-Forget. Mixed Approaches to Named Entity Recognition*. Proceedings of 6th International Workshop on Applications of Natural Language for Information Systems. GI-Edition. Lecture Notes in Informatics. vol. 3. Madrid: 2001.

K.L. Wong, W. Lam, J. Yen, *Interactive Chinese News Event Detection and Tracking*, Proceedings of The Second Asian Digital Library Conference, 1999, pp.30-43.

Y. Yang, T. Pierce, J. Carbonell, *A Study on Retrospective and On-Line Event Detection*, Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval, 1998, pp.28-36.