Advanced Clustering Technique for Medical Data Using Semantic Information*

Kwangcheol Shin[1], Sang-Yong Han[1], and Alexander Gelbukh[1,2]

[1] School of Computer Science and Engineering, Chung-Ang University,
221 HukSuk-Dong, DongJak-Ku, Seoul, 156-756, Korea
kcshin@archi.cse.cau.ac.kr, hansy@cau.ac.kr
[2] Center for Computing Research,
National Polytechnic Institute, Mexico City, Mexico
{gelbukh,igor}@cic.ipn.mx, www.Gelbukh.com

Abstract. MEDLINE is a representative collection of medical documents supplied with original full-text natural-language abstracts as well as with representative keywords (called MeSH-terms) manually selected by the expert annotators from a pre-defined ontology and structured according to their relation to the document. We show how the structured manually assigned semantic descriptions can be combined with the original full-text abstracts to improve quality of clustering the documents into a small number of clusters. As a baseline, we compare our results with clustering using only abstracts or only MeSH-terms. Our experiments show 36% to 47% higher cluster coherence, as well as more refined keywords for the produced clusters.

1 Introduction

As science and technology continues to advance, the number of related documents rapidly increases. Today, the amount of related stored documents is incredibly massive. Information retrieval on the numerous documents has become an active field for study and research.

MEDLINE database maintained by the National Library of Medicine (http://www.nlm.gov/) contains ca. 12 million abstracts on biology and medicine collected from 4,600 international biomedical journals, stored and managed in the format of eXtensible Markup Language (XML). MeSH (Medical Subject Headings) are manually added to each abstract to describe its content for indexing. Currently, MEDLINE supports various types of search queries.

However, query-based information search is quite limited on performing searches for biological abstracts. The query-based search method may be appropriate for content-focused querying, but this is so only on the condition that the user is an expert in the subject and can choose the keywords for the items they are searching for. Thus it is very confusing and time-consuming for users who are not experts in biology or

* Work done under partial support of the ITRI of Chung-Ang University, Korean Government (KIPA Professorship for Visiting Faculty Positions in Korea), and Mexican Government (CONACyT, SNI, IPN). The third author is currently on Sabbatical leave at Chung-Ang University.

R. Monroy et al. (Eds.): MICAI 2004, LNAI 2972, pp. 322–331, 2004.

medicine to search using queries because new techniques and theories pour out continuously in this particular field. Even expert users have troubles when they need to quickly find specific information because it is impossible to read every found document from the beginning to the end. We see that query based search method is not so efficient in both cases [1].

Unlike the general query-based search method, document clustering automatically gathers highly related documents into groups. It is a very important technique now that efficient searching must be applied to massive amounts of documents. Generally, unsupervised machine learning [2] method is applied for clustering. Then the features included in the instance are provided as input of the clustering algorithm in order to group together highly similar documents into a cluster.

In recent years, concerns on processing biological documents have increased but were mostly focused on the detection and extraction of relations [3][4] and the detection of keywords [5][6]. Up to now, few studies on clustering of biological documents have been done, except for the development of TextQuest [1], which uses the method of reducing the number of terms based on the abstract of the document. Namely, this is simple method uses a cut-off threshold to eliminate infrequent terms and then feeds the result into an existing clustering algorithm.

The present study proposes an improvement to clustering technique using the semantic information of the MEDLINE documents, expressed in XML. For the proposed method the XML tags are used to extract the so-called MeSH, the terms that represent the document, and give additional term weights to the MeSH terms in the input of a clustering algorithm. The greater the additional term weight given to the MeSH, the greater the role of the MeSH terms in forming clusters. This way the influence of the MeSH on the clusters can be controlled by adjusting the value of the additional term weighting parameter.

For this study, the vector space model [7] with the cosine measure was used to calculate the similarity between the documents. The formula for calculating the coherence of the cluster was applied to evaluate the quality of the formed clusters, and the cluster key words were extracted based on the concept vector [8] for summarizing the cluster's contents.

2 Basic Concepts

This section will present the theoretical background: the vector space model, which is the basis for document searching, the measurement of coherence for evaluating the quality of the clusters, and the extraction of keywords that represent the cluster.

2.1 Vector Space Model

The main idea of the vector space model [7] is expressing the documents by the vector of its term frequency with weights. The following procedure is used for expressing documents by vectors [9]:

- Extract all the terms from the entire document collection.
- Exclude the terms without semantic meaning (called stopwords).

- Calculate the frequency of terms for each document.
- Treat terms with very high or low frequency as stopwords.
- Allocate indexes from 1 to d to each remaining term, where d is the number of such terms. Allocate indexes from 1 to n to each document. Then the vector space model for the entire group of documents is determined by the $d \times n$-dimensional matrix $w = | w_{ij} |$, where w_{ij} refers to the *tf-idf* (term frequency—inverse document frequency) value of the i-th term in j-th document.

Measuring the similarity between two documents in the process of clustering is as important as the selection of the clustering algorithm. Here we measure the similarity between two documents by the cosine expression widely used in information retrieval and text mining, because it is easy to understand and the calculation for sparse vectors is very simple [9].

In this study, each of the document vector $x_1,x_2,\ldots,x_n$ was normalized in the unit of the L_2 norm. The normalization here only maintains the terms' direction, to have the documents with the same subject (i.e., those composed of similar terms) converted to similar document vectors. With this, the cosine similarity between the document vectors x_i and x_j can be derived by using the inner product between the two vectors:

$$s(x_i, x_j) = x_i^T x_j = \| x_i \| \| x_j \| \cos(\theta(x_i, x_j)) = \cos(\theta(x_i, x_j))$$

The angle formed by the two vectors is $0 \le \theta(x_i,x_j) \le \pi/2$.

When n document vectors are distributed into k disjoint clusters $\pi_1,\pi_2,\ldots,\pi_k$, the mean vector, or centroid, of the cluster π_j is

$$m_j = \frac{1}{|\pi_j|} \sum_{x \in \pi_j} x \tag{1}$$

Then, when the mean vector m_j is normalized to have a unit norm, the concept vector $c_{j,}$ can be defined to possess the direction of the mean vector [8].

$$c_j = \frac{m_j}{\|m\|} \tag{2}$$

The concept vector c_j defined in this way has some important properties, such as the Cauchy-Schwarz inequality applicable to any unit vector z:

$$\sum_{x \in \pi_j} x^T z \le \sum_{x \in \pi_j} x^T c_j \tag{3}$$

Referring to the inequality above, it can be perceived that the concept vector c_j has the closest cosine similarity with all document vectors belonging to cluster π_j.

2.2 Cluster Quality Evaluation Model

As a measure of quality of obtained clusters we use the coherence of a cluster π_j ($1 \le j \le k$), which based on formula (3) can be measured as

$$\frac{1}{|\pi_j|} \sum_{x \in \pi_j} x^T c_j \tag{4}$$

If the vectors of all the documents in a cluster are equal, its coherence is 1, the highest possible value. On the other hand, if the document vectors in a cluster are spread extremely wide apart from each other, the average coherence would be close to 0. The coherence of cluster system $\pi_1,\pi_2,\ldots,\pi_k$ is measured using the objective function shown below [8]:

$$Q(\{\pi_j\}_{j=1}^{k}) = \sum_{j=1}^{k} \sum_{x_i \in \pi_j} x_i^T c_j \qquad (5)$$

The coherence measure can be successfully used when the number of clusters is fixed, as it is in our case (with a variable number of clusters it favors a large number of small clusters).

2.3 Extracting Cluster Summaries

To present the clusters to the user, we provide cluster summaries, for the user to understand easier the contents of the documents in the cluster and to be able to select the cluster of interest by examining only these summaries. As summaries we use representative keyword sets.

Given n document vectors divided into k disjoint clusters $\pi_1,\pi_2,\ldots,\pi_k$, denote the keywords of the document cluster π_j by $words_j$. A term is included in the summary $word_j$ if its term weight in the concept vector c_j of cluster π_j is greater that its term weight in the concept vectors of other clusters [8]:

$$words_j = \{k^{\text{th}} \text{ word: } 1 \le k \le d,\ c_{k',j} \ge c_{k',m},\ 1 \le m \le c,\ m \ne j\}, \qquad (6)$$

where d is the total number of terms and $c_{k',j}$ is the weight of the j-th term of k-th cluster. Such summary $word_j$ has the property of having the concept vector localized to the matching cluster. Because of this, it gives comparably good keywords.

3 Term Weighing

This section presents our method of advanced clustering technique combining the structured semantic information assigned to the MEDLINE documents with their full-text information.

3.1 Medical Subject Headings (MeSH)

MeSH is an extensive list of medical terminology. It has a well-formed hierarchical structure. MeSH includes major categories such as *anatomy/body systems*, *organisms*, *diseases*, *chemicals and drugs* and *medical equipment*. Expert annotators of the National Library of Medicine databases, based on indexed content of documents, assign subject headings to each document for the users to be able to effectively retrieve the information that explains the same concept with different terminology.

MeSH terms are subdivided into Major MeSH headings and MeSH headings. Major MeSH headings are used to describe the primary content of the document, while MeSH headings are used to describe its secondary content. On average, 5 to 15 subject headings are assigned per document, 3 to 4 of them being major headings.

MeSH annotation scheme also uses subheadings that limit or qualify a subject heading. Subheadings can also be major subheadings and subheadings according to the importance of the corresponding terms.

3.2 Extraction of MeSH Terms from MEDLINE

MEDLINE annotated documents are represented as XML documents. An XML document itself is capable of supporting tags with meaning. The MEDLINE documents are formed with Document Type Description (DTD) that works as the document schema, namely, NLM MEDLINE DTD (www.nlm.nih.gov/database/dtd/nlmcommon.dtd). In the DTD, information on the MeSH is expressed with the <MeshHeadingList> tag as shown in Fig. 1. Fig. 2 shows the MeSH expressed by using the DTD of Fig. 1.

```
<!ELEMENT MeshHeadingList (MeshHeading+)>
<!ELEMENT MeshHeading (Descriptor, QualifierName* )>
<!ELEMENT Descriptor (#PCDATA)>
<!ATTLIST Descriptor MajorTopicYN (Y | N) "N">
<!ELEMENT QualifierName (#PCDATA)>
<!ATTLIST QualifierName MajorTopicYN (Y | N) "N">
```

Fig. 1. NLM MEDLINE DTD (MeSH)

```
<MeshHeadingList>
<MeshHeading><Descriptor MajorTopicYN="Y">Acetabulum</Descriptor></MeshHeading>
<MeshHeading><Descriptor MajorTopicYN="N">Adolescence</Descriptor></MeshHeading>
<MeshHeading><Descriptor MajorTopicYN="N">Adult</Descriptor></MeshHeading>
<MeshHeading><Descriptor MajorTopicYN="N">Aged</Descriptor></MeshHeading>
<MeshHeading><Descriptor MajorTopicYN="N">Arthroscopy</Descriptor></MeshHeading>
<MeshHeading><Descriptor MajorTopicYN="N">Cartilage, Articular</Descriptor>
        <QualifierName MajorTopicYN="Y">injuries</QualifierName>
        <QualifierName MajorTopicYN="N">surgery</QualifierName></MeshHeading>
<MeshHeading><Descriptor MajorTopicYN="N">Case Report</Descriptor></MeshHeading>
<MeshHeading><Descriptor MajorTopicYN="N">Female</Descriptor></MeshHeading>
<MeshHeading><Descriptor MajorTopicYN="N"">Human</Descriptor></MeshHeading>
<MeshHeading><Descriptor MajorTopicYN="N">Male</Descriptor></MeshHeading>
<MeshHeading><Descriptor>Rupture</Descriptor></MeshHeading>
<MeshHeading><Descriptor MajorTopicYN="Y">Surgical Procedures, Endoscopic</Descriptor>
</MeshHeading>
</MeshHeadingList>
```

Fig. 2. MEDLINE Data Sample (MeSH)

As observed in Fig. 1 and Fig. 2, the information on MeSH in DTD can be classified largely into two types, which can in turn be divided into two more specific types each. The first type refers to the terms appearing under the tag <*Descriptor*> with *MajorTopicYN* attribute *Y* or *N*, which indicates that the terms appearing under the tag are major MeSH or MeSH, accordingly. The second type refers to terms appearing under the tag <*QualifierName*> with the *MajorTopicYN* attribute *Y* or *N*, which indicates that the terms appearing under the tag are major subheadings or subheadings.

In order to extract the four types of terms described above, the XML document is parsed and the DOM (www.w3.org/DOM) tree is produced. The terms extracted in this process are used to adjust the term weights in the document vector.

3.3 Combining Abstract Terms with MeSH Terms

MEDLINE documents include both full-text abstract taken from the original document (a scientific article published in a journal) as well as the MeSH terms. The MeSH terms represent important information about the document, since they are assigned by experts with great care and consideration. Thus it can be expected (and our experiments confirmed this) that clustering the documents according to their MeSH information would give better clusters than clustering using the original abstracts.

However, combining the two types of information can lead to even better results. We combine the two sources of information by adjusting term weights in the document vectors used as feature weight by the clustering algorithm.

Indeed, clustering results are highly influenced by the term weights assigned to individual terms; for better results, more important terms are to be assigned greater weight. Since MeSH terms are known to be more important, the original term weights w_{ij} for the words appearing under the corresponding headings are adjusted as follows:

$$w_{ij} \leftarrow \begin{cases} w_{ij} + (\rho + \dfrac{\rho}{4+\ln(\rho)}); & <Decriptor\ MajorTopicYN = "Y"> \\ w_{ij} + (\rho + \dfrac{\rho}{2\times(4+\ln(\rho))}); & <Decriptor\ MajorTopicYN = "N"> \\ w_{ij} + (\rho - \dfrac{\rho}{2\times(4+\ln(\rho))}); & <SubHeading\ MajorTopicYN = "Y"> \\ w_{ij} + (\rho - \dfrac{\rho}{4+\ln(\rho)}); & <SubHeading\ MajorTopicYN = "N"> \end{cases} \quad (7)$$

In this formula, a coefficient ρ is used to control de degree in which the abstracts or the MeSH terms participate in the resulting weighting. With ρ close to 0, the MeSH terms do not receive any special treatment, so that the results are close to those of clustering using only abstracts. With very large ρ, the results are close to those of clustering using only MeSH terms. With intermediate values of ρ, the two types of information are combined. Our experimental results show that the best clusters are obtained with some intermediate values of ρ.

The specific expressions in formula (7) were found empirically in such a way that the formula gives slightly different additional values to the terms according to their significance: about 33% and 16% of the value of ρ is added or subtracted from its original value. For example, when $\rho = 0.3$, additional term weights are $0.3 + 0.107$, $0.3 + 0.054$, $0.3 – 0.054$, and $0.3 – 0.107$, respectively.

After the term weights are modulated by the above formula, they are re-normalized since the former normalized value had been changed, see Fig. 3.

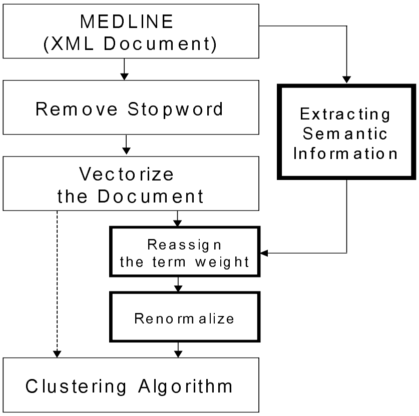

Fig. 3. The Main Algorithm.

4 Experimental Results

For the experiment, 4,872 documents have been extracted from MEDLINE edition published in 1996. Two groups of documents were formed: one group contained documents with abstract only and another one those with both abstract and MeSH terms.

The MC program [10], which produces vectors from a given group of documents, was used to vectorize the documents. Stopwords and the terms with frequency lower than 0.5% and higher than 15% were excluded. With this, the document group with abstracts only had 2,792 terms remaining and that with both abstract and MeSH, 3,021. Then the *tf-idf* value was calculated for each of the document groups and normalized to the L_2 norm to form 4,872 document vectors.

To verify the proposed method, the MeSH terms were extracted from the document group with both abstract and MeSH and then for the extracted terms, the term weights of the corresponding terms of each document vector were modulated by (3-1). Then they were normalized to the L_2 norm.

The standard spherical k-means algorithm was implemented for testing these data. This is an efficient clustering algorithm that quickly produces the fixed number of clusters specified by the user.

We clustered our test document set into a fixed number of 3 to 6 clusters (these numbers are of major interest for user interfaces providing document collection navigation support) and using different values of the parameter ρ, varying smoothly from abstract-only to MeSH-only strategies. The abstract-only strategy was used as a baseline. The quality of the clusters was measured as inter-cluster coherence, as explained in Section 2.2. The test results are as shown in Table 1, where the gain ratio is calculated relative to the abstract-only strategy. The best values in each column are emphasized. Fig. 4 shows the average (over different number of clusters; the right-most column of Table 1) coherence obtained with different values of the parameter ρ.

As can be observed in the figure, there is a wide area of the values of ρ providing optimal values of coherence. In particular, such optimal values are 36 to 47% better than those obtained with abstracts only and 10 to 15% better than with MeSH terms only. This justifies our idea of combination of these two sources of information in a non-trivial manner.

Table 1. Experimental Results.

N stands for the number of clusters, C for average coherence of the obtained clusters, and R for gain rate relative to the abstract-only clustering.

ρ	N = 3		N = 4		N = 5		N = 6		Average	
	C	R	C	R	C	R	C	R	C	R
Abstracts only	1065	0%	1126	0%	1157	0%	1234	0%	1146	0%
ρ = 0	1185	11.3%	1217	8.1%	1280	10.6%	1327	7.5%	1252	9.3%
ρ = 1	1480	39.0%	1561	38.6%	1619	39.9%	1679	36.1%	1585	38.3%
ρ = 5	1535	44.1%	1560	38.5%	1650	42.6%	1717	39.1%	1616	41.0%
ρ = 10	1557	46.2%	1631	44.8%	1684	45.5%	1720	39.4%	1648	43.8%
ρ = 20	1560	46.5%	1616	43.5%	1692	46.2%	1716	39.1%	1646	43.7%
ρ = 100	**1570**	47.4%	**1641**	45.7%	**1699**	46.8%	**1746**	**41.5%**	**1664**	45.3%
ρ = 500	1550	45.5%	1619	43.8%	1685	45.6%	1743	41.2%	1649	44.0%
ρ = 1000	1558	46.3%	1585	40.8%	1664	43.8%	1738	40.8%	1636	42.8%
ρ = 2000	1563	46.8%	1601	42.2%	1688	45.9%	1655	34.1%	1627	42.0%
MeSH only	1361	27.8%	1422	26.3%	1535	32.7%	1561	26.5%	1470	28.3%

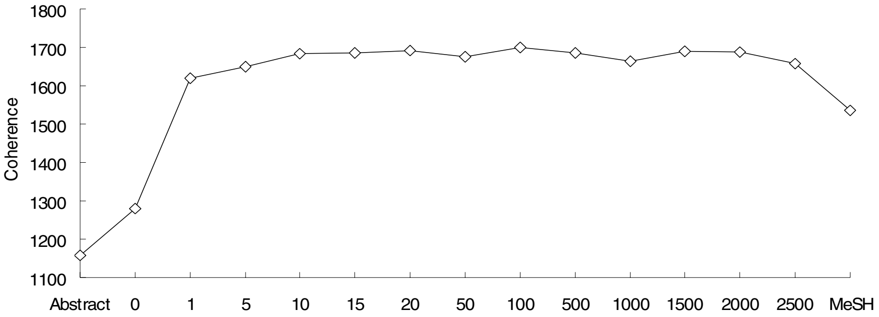

Fig. 4. Experimental Results: Average coherence as function of the parameter ρ.

As an additional evidence of the improvement in cluster quality, we extracted the keyword summaries from the clusters produced with the baseline procedure and with our method, accordingly. For this, we used the formula (6). Table 2 represents the summaries for the five clusters obtained using abstracts only and Table 3 for the clusters obtained with $\rho = 100$, which gives the best clusters according to the coherence measure. Fifteen key words with the greatest term weight were extracted from the five clusters.

One can easily observe that the keyword summary is more consistent and natural for the clusters obtained with a non-trivial value of the parameter ρ.

Table 2. Keyword summaries for clustering based only on abstracts.

Cluster	Key Words
1	*cells, protein, cell, dna, proteins, expression, receptor, gene, beta, binding, alpha, human, acid, activity, kinase*
2	*care, health, medical, united, states, usa, medicine, apr, management, clin, research, cancer, nursing, dis, nurse*
3	*hiv, mice, virus, peptide, infected, california, muscle, model, signaling, mucosal, wild, differentiation, class, produced, signal*
4	*patients, group, treatment, clinical, cases, trial, study, patient, disease, surgery, risk, years, age, hospital, children*
5	*heart, coronary, rats, cardiac, ventricular, cardiology, artery, aim, myocardial, pressure, failure, atrial, flow, exercise, blood*

Table 3. Keyword summaries for clustering by abstracts and MeSH terms (ρ = 100).

Cluster	Key Words
1	*cell, cells, mice, pathology, expression, growth, factor, tumor, cultured, kinase, induced, antigens, hiv, membrane, beta*
2	*human, health, care, united, states, medical, nursing, drug, agents, medicine, disease, research, patient, therapy, hospital*
3	*diagnosis, female, male, age, case, adult, middle, aged, radiography, diseases, heart, child, neoplasms, ultrasonography, adolescence*
4	*animal, rats, support, receptors, chemical, effects, brain, activity, wistar, sprague, dawley, inhibitors, antagonists, muscle, rat*
5	*sequence, dna, proteins, protein, molecular, acid, amino, binding, data, gene, structure, base, genetic, recombinant, genes*

5 Conclusion

Medical documents in the MEDLINE database contain both original full-text natural-language abstract and structured keywords manually assigned by expert annotators. We have shown that a combination of these two sources of information provides better features for clustering the documents than each of the two sources independently. We have shown a possible way of their combination, depending on a parameter that determines the degree of contribution of each source. Namely, we combined them by adjusting the term weights of the corresponding terms in the document vectors, which were then used as features by a standard clustering algorithm.

Our experimental results show that there exists a wide area of the values of the parameter that gives a stable improvement in the quality of the obtained clusters, both in internal coherence of the obtained clusters and in the consistency of their keyword summaries. This implies, in particular, that our method is not sensible to a specific selection of the parameter.

The improvement observed in the experiments was 36 to 47% in comparison with taking into account abstracts only and 10 to 15% in comparison with taking into account MeSH terms only.

References

[1] Iliopoulos I, Enright A, Ouzounis C.: Textquest: document clustering of medline abstracts for concept discovery in molecular biology. Pac. Symp. on Biocomput. 384-395. 2001.

[2] Kubat M., Bratko I. and Michalski R.S., In Machine Learning and Data Mining: methods and applications,: A review of machine learning methods, Ed. Michalski R.S., Bratko I. and Kubat M., John Wiley & Sons, New York NY, 1997.

[3] Sekimizu T., Park H. S. and Tsujii J.,: Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts, Genome Informatics Workshop, Tokyo, p. 62, 1998.

[4] Thomas J., Milward D., Ouzounis C., Pulman S. and Carroll M.,: Automatic extraction of protein interactions from scientific abstracts, Pac. Symp. Biocomput, p. 538-549, 2000.

[5] Andrade M.A. and Valencia A.,: Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families, Bioinformatics 14, 600, 1998.

[6] Proux D., Rechenmann F., Julliard L., Pillet V. and Jacq B.,: Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction, Genome Informatics Workshop, Tokyo, p.72-80, 1998.

[7] Salton G. and. McGill M. J.,: Introduction to Modern Retrieval, McGraw-Hill Book Company, 1983.

[8] Dhillon I. S. and Modha, D. S.: Concept Decomposition for Large Sparse Text Data using Clustering, Technical Report RJ 10147(9502), IBM Almaden Research Center, 1999.

[9] Frakes W. B. and Baeza-Yates R.: Information Retrieval : Data Structures and Algorithms, Prentince Hall, Englewood Cliffs, New Jersey, 1992.

[10] Dhillon I. S., Fan J., and Guan Y.,: Efficient Clustering of Very Large Document Collections. Data Mining for Scientific and Engineering Applications, Kluwer Academic Publishers, 2001.