Selection of Representative Documents for Clusters in a Document Collection*

Alexander Gelbukh [1], Mikhail Alexandrov [1], Ales Bourek [2], Pavel Makagonov [3]

[1] Center for Computing Research (CIC), National Polytechnic Institute (IPN),
Av. Juan de Dios Batiz, C.P. 07738, D.F., Mexico
{gelbukh, dyner}@cic.ipn.mx, www.gelbukh.com

[2] Faculty of Medicine, Masaryk University in Brno,
Jostova 10, 66243 Brno, Czech Republic
bourek@med.muni.cz

[3] Mixteca University of Technology,
Km 2.5 Carretera Acatlima, Huajuapan de Leon, Oaxaca, C.P. 69000, Mexico
mpp2003@inbox.ru

Abstract. An efficient way to explore a large document collection (e.g., the search results returned by a search engine) is to subdivide it into clusters of relatively similar documents, to get a general view of the collection and select its parts of particular interest. A way of presenting the clusters to the user is selection of a document in each cluster. For different purposes this can be done in different ways. We consider three cases: selection of the average, the "most typical," and the "least typical" document. The algorithms are given, which rely on a dictionary of keywords reflecting the topic of the user's interest. After clustering, we select a document in each cluster basing on its closeness to the other ones. Different distance measures are discussed; preliminary experimental results are presented. Our approach was implemented in the new version of Document Classifier system.

1. Introduction

In many cases one needs to divide the set of the documents into smaller groups and, instead of considering the whole group, to choose a representative element in each group for closer examination.

Selection of a set of documents that, in some sense, are representative for a given collection has been considered in the literature without relation to subdivision of the collection into clusters: such documents represent the collection as a whole and not its specific subsets [5]. For example, such representative documents can be those that are most closely relater to a given thematic domain. A popular technique used for this purpose is Bayesian classifiers [9]. Even when the selected elements are intended to represent specific clusters, usually simply the centers of the clusters are considered [3][4][8].

* Work done under partial support of Mexican Government (CONACyT, SNI) and CGPI-IPN.

However, in different situations different elements should represent a cluster, e.g:

- Case A: the "typical" (*average*) element is gives the idea of its group. Consider a specialist planning future research on a specific problem using a digital library. The first task is to identify various sub-domains, or aspects, of the whole problem. It is helpful to automatically cluster all papers on the given problem into several groups and figure out what each group is about. For this, one can read the *typical* paper automatically selected by our program in each cluster. This allows selecting the most interesting cluster for more detailed reading.
- Case B: the "*least typical*" element is good for achieving agreement. To organize a discussion between specialists that have submitted proposals on a certain problem, one needs to discover what (groups of) opinions there are and select a representative of each such group for a forum where consensus is to be achieved. Thus, the representative element (the author of the selected proposal) not only is to belong to his or her group but also should be the most familiar with the other points of view. Thus, the organizer can cluster the proposals and automatically select in each cluster the one most similar to the rest of the collection.
- Case C: the "*most typical*" element gives the idea of the differences. In a Chinese restaurant one is offered the "typical Chinese" food. This is not the average (over all Chinese people and all days of year) food that the Chinese eat (that would be rice) but the "least European" food that they eat. Similarly, the "typical" (i.e., less Western) Russian wear is sarafan while the "average" would be jeans. This kind of "the most typical" element is good to illustrate the diversity and emphasize the differences between the groups. In a set of documents, one can be interested in reading the ones that show the diversity of the clusters without much intersection.

In the case A the selected document is the nearest one to all other documents in its cluster (this is usually referred to as the *centroid* of the cluster). In the other two examples the selected document is at the "border" of the cluster and not in its "center," see Figure 1.

In this paper we present the algorithms that, after clustering the document collection, find the corresponding representative elements in each cluster. Specifically, we refer to the implementation of the corresponding algorithms in the new version of the system Document Classifier for interactive computer-aided exploration and classification of large document collections.

2. Numerical Representation of the Documents

2.1. Document Image

A document image is a numerical vector corresponding to the density of keywords from a domain dictionary [7]. A domain dictionary (DD) is a list of keywords w_k supplied with the coefficients A_k of their importance for the given topic (thematic domain). These coefficients are numbers between 0 and 1 that reflect the fuzzy nature of the relationship between the keywords and the selected domain. Let n_k be the number of occurrences of

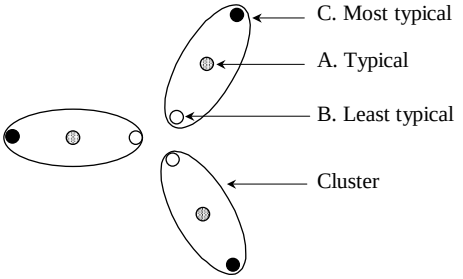

Figure 1: Types of representative elements.

the keyword w_k in the document. Then the document image is the vector (X_1, X_2, ..., X_N) such that $X_k = A_k n_k$, where A_k are the coefficients of importance of w_k in the DD.

The direction of this vector (here we ignore its length) reflects the *theme of the document*. Indeed: a document consisting of several concatenated copies of the same document—which, obviously, is devoted to the same theme—has the same direction of the document image vector, even if differs in length.

We will also need two other vector representations of the document that can be obtained from the vector (X_k). One is the *binary document image* (I_1, I_2, ..., I_N), where $I_k = 0$ if $X_k = 0$ and $I_k = 1$ otherwise.

The other is the *normalized document image* (x_1, x_2, ..., x_N), where $x_k = X_k / M$; here M is a number of running words in the document (including keywords, but usually excluding stop-words like prepositions, etc.). This normalization means that we reduce all estimations to the average per word.

2.2. Closeness between a Document and a Given Topic

For evaluation of the closeness between a document and the domain we use coverage of the DD's keyword list by the document and, correspondingly, the coverage of the document by the DD:

$$W_l = \frac{\sum_k A_k I_k}{\sum_k A_k}; \quad W_d = \sum_k x_k \tag{1}$$

The measure W_l reflects the density in the DD of the words that occur in the document; it does not depend on the scale of the coefficients A_k. The measure W_d reflects the density of the keywords in the document and does not depend on the size of the document. Both values belong to the interval [0,1]; each of them can be considered as the degree of closeness between the document and the domain.

These measures give different kinds of information on the closeness between the document and the domain. If a small set of certain keywords is repeated many times in the document, the W_l is low even though the W_d is high. If most keywords do occur in the document but their occurrences are very rare then W_l is high, but W_d is low. In both cases the given topic is not represented well in the document; thus, for such similarity both measures should be high.

Accordingly, though in our program the user can use each measure separately, it is often desirable to take into account both measures simultaneously. This can be achieved by their combination:

$$W = \alpha W_l + \beta W_d, \quad \alpha + \beta = 1. \tag{2}$$

Here, α and β are user-defined weights. Usually in practice we set $\alpha = 0.05$ and $\beta = 0.95$, because W_d does not exceed 5% for almost all practically interesting cases. It means that on average one keyword occurs in no more than 3 sentences in domain-oriented documents (given that the average sentence length in European languages is about 7 words).

2.3. Closeness between Documents

Let $(x_{11}, x_{21}, ..., x_{N1})$ and $(x_{12}, x_{22}, ..., x_{N2})$ be the normalized images of two documents. To estimate the distance D between them the well-known cosine and polynomial measures of various degrees are used:

$$D_c = 1 - R, \quad D_p = \sqrt[p]{\sum_k (x_{k1} - x_{k2})^p}\ . \tag{3}$$

Here R is the normalized correlation coefficient:

$$R = \frac{\sum_k (x_{k1}\, x_{k2})}{\|x_1\| \|x_2\|}, \quad \|x_i\| = \sqrt{\sum_k x_{ki}^2}$$

and $p = 1, 2, 4, ..., \infty$; the case $p = \infty$ corresponds to $D_p = \max_k |x_{k1} - x_{k2}|$.

Again, these measures reflect different aspects of similarity and are to be used in different situations. The cosine measure is preferable if the user wants to compare the themes of the two documents. If the user wants to compare the coverage of the documents by the domain, then the polynomial measure is to be used. In the latter case, by increasing the degree p the user can emphasize large differences in the numbers of occurrences of few keywords in the two documents.

In practice it is often desirable to combine both considerations, i.e., to compare both the themes and coverage levels of two documents. For this, the two measures are combined:

$$D = \gamma D_c + \delta D'_p, \quad \gamma + \delta = 1. \tag{4}$$

Here D'_p is proportional to D_p, see (5) below; γ and δ are user-defined weights—the penalties for the difference in the themes and in the coverage, correspondingly. In practice we set $\gamma = \delta = 0.5$. Combination of the measures is discussed in [1].

When combining different measures, it is convenient to scale them to the same interval; this only affects the choice of the weights γ and δ. Since the cosine measure varies in the interval [0, 1], we scale the polynomial measure accordingly:

$$D'_p = \frac{D_p}{\max D_p} = \frac{D_p}{\max\limits_k A_k}, \tag{5}$$

since the maximum possible value of D_p is the maximal coefficient A_k.

Given a set of N documents and a measure, e.g., (4) for calculating the distance D_{ik} between the *documents* i and k, we can calculate the average distance between a given document i and other documents $k \neq i$ in the set:

$$D_i = \frac{1}{N} \sum_{k:\, k \neq i} D_{ik} \tag{6}$$

3. Choice of Representative Documents

3.1. Clustering

It is well known that the number of methods and their modifications used in cluster analysis is greater than the number of authors working in this area. For simplicity, we implemented in our system only two methods: the simplest hierarchical method (the *nearest neighbor* method) and the simplest non-hierarchical method (*K-means*).

The former method builds a dendrite and then eliminates the weak connections so that instead of one tree several sub-trees appear. Each sub-tree is considered a cluster reflecting a specific sub-domain. In the latter method, the desired number of clusters is set by the user. There is extensive literature discussing such methods and their applications in text processing [4][7]. Discussion of the clustering methods we use is beyond the scope of this paper.

3.2. Choice of the Representative Element in a Cluster

After the collection has been subdivided into clusters, a representative element can be chosen in each cluster according to the task under consideration. This element represents its cluster in various situations where only one member of each cluster should be selected; for different tasks different representatives are to be chosen—see examples in Section 1. Accordingly, different criteria for the choice of the representative document in a cluster can be suggested, which correspond to the examples from Section 1:

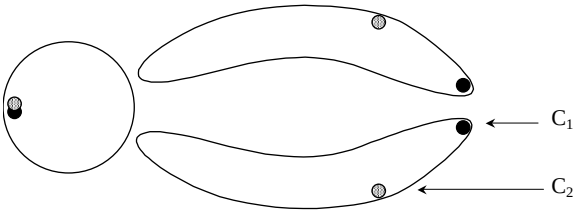

Figure 2: Types of the "*most typical*" elements.

- Case A: *Maximum closeness to the other documents in the cluster*. The document with the minimum average distance D_i from the other documents in the cluster is chosen; D_i is given by (6) with summation by the documents of the given cluster.
- Case B: *Maximum closeness to the domain*. The document with the maximum value given by (2) is chosen.
- Case C_1: *Maximum distance from the domain*. Unlike Case B, the document with the *minimum* value given by (2) is chosen.
- Case C_2: *Maximum distance from the documents in the other clusters*. Unlike Case A, the document is chosen with the *maximum* value D_i given (6) with summation over all documents of the whole collection *except* the ones belonging to the given cluster.

The difference between the results obtained for the cases A, B, and C is illustrated in Figure 1. In the case C, there are two possible variants of the task (in fact, two different tasks), hence two different algorithms; the difference is illustrated in Figure 2.

In our program, the user can select one of these criteria according to a specific task. E.g., the criterion A can be used in the situation described in the example A, Section 1—the choice of the texts to read. Indeed, they must reflect the contents of all texts in their groups. Here, it is not so important how close the selected documents are to the global domain under consideration. On the other hand, in the example B from Section 1—selection of representatives for negotiations—the persons familiar with the problem in the whole should be chosen in each group; in this case the criterion B is to be used.

4 A Practical Example

We have analyzed papers from a medical domain. Selection and classification of medical documents is very important for the experts of Czech Ministry of Healthcare, so that high qualitative classifiers are used in their everyday activity [2].

The document collection contained 98 papers. We knew in advance that there were papers on cardiology, urology, therapeutics, etc., in total 8 sub-domains. First, we constructed the DD for all documents as described in [6]. Then we clustered the documents using the nearest neighbor algorithm. We tried various thresholds for the admissible distance between documents. Only with the cosine measure we obtained the expected number of clusters (8); this means that the documents contained much information noise that was

filtered out by the cosine measure. For the selection of representative documents, we chose the criterion of maximum closeness to the other elements of the cluster (case A) and the closeness to the domain (case B).

The results of clastering matched the opinion of a human expert in 84% of the documents. The representative documents of the clusters selected by the program were within 3 best candidates selected by the human expert, for 5 clusters when calculated by the scheme A and for all clusters by the scheme B. This means that the expert (intuitively) selected them according to the case B and not A.

5 Conclusions

A domain dictionary consisting of domain-specific keywords gives a possibility to build various numerical measures for evaluation of closeness between documents and between a document and a given domain. A combination of measures allows taking into account the user preferences and the specific task the user is to accomplish.

To represent a cluster, we choose one of its elements. For different goals, different strategies for choosing such elements are used. We have discussed the corresponding algorithms and the quantitative characteristics of documents and domains used by them.

References

[1] Alexandrov, M., Gelbukh, A., and Makagonov, P. (2000): *On metrics for Keyword-Based document selection and classification.* In: A. Gelbukh (Ed.) Proc. of CICLing-2000, 1st Intern. Conf. on Intelligent Text Processing and Comp. Linguistics, Mexico, pp. 373–389.

[2] Bourek, A. (2002): *The Era of Information in the Czech Republic: How Healthcare Is Managing Data Sets and Mind-Sets*. In: Vahe A. Kazandjian (Ed.). Accountability through measurement: a global healthcare imperative. American Society of Quality, Milwaukee, Wisconsin, ASQ Quality Press, pp.291-324.

[3] Chaudhur, D., Murthy C., Chaudhur B. (1994): *Finding a Subset of Representative Points in a Data Set*. IEEE Trans. on Systems, Man, and Cybernetics, N 24 (9), pp.1416–1424.

[4] Hartigan, J. (1975): *Clustering Algorithms.* Wiley.

[5] Kreuzman, H. (2001): *A Cocitation Analysis of Representative Authors in Philosophy-Examining the Relationship between Epistemologists and Philosophers of Science.* Scientometrics, N 51 (3), pp.525-539.

[6] Makagonov, P., Alexandrov, M., Sboychakov, K. (2000): *A toolkit for development of the domain-oriented dictionaries for structuring document flows*. In: Kiers H.A. et al. (Eds.) Data Analysis, Classification, and Related Methods, Springer, pp.83-88

[7] Manning, D.C., Schutze, H. (1999): *Foundations of statistical natural language processing*. MIT Press.

[8] Moens, M., Uyttenda, C., Dumortie, J. (1999): *Abstracting of Legal Cases – The Potential of Clustering based on the Selection of Representative Objects.* Journal of the Amer. Soc. for Inform, Science, N 50 (2), pp. 151-161.

[9] Zizka, J., Bourek, A. (2002): *Automated Selection of Interesting Medical Text Documents by the TEA Text Analyzer*. In: A. Gelbukh (Ed.) Comp. Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science N 2276, Springer-Verlag, pp. 402–404.