Selection of Representative Documents in a Document Collection *

PAVEL MAKAGONOV [1], MIKHAIL ALEXANDROV [2], ALEXANDER GELBUKH [2],

[1] Moscow Mayor's Directorate, Moscow City Government,
Novij Arbat 36, Moscow, 121205, RUSSIA

[2] Center for Computing Research (CIC), National Polytechnic Institute (IPN),
Av. Juan de Dios Batiz, C.P. 07738, D.F., MEXICO

Abstract: - In different situations, different documents can be selected as representative ones of document groups found in a large document set. We consider three different problems of this kind—selection of the average document, the "most typical" document and the "least typical" one, giving the corresponding algorithms. These tasks are considered in the framework of a given topic defined by a domain-oriented keyword dictionary. The procedure consists of two phases: (1) clustering documents into sub-topics and (2) definition of the representative document in each group. For the latter, the notions of *potential* and *difference of potentials* are introduced, which are applied to the dendrite constructed by the method of the nearest neighbor. Unlike the traditional clustering on dendrite, the potentials allow to take into account the structure of connections in significantly greater detail. Por approach has been implemented in a new version of the system *Text Classifier*.

Key-Words: - Natural Language Processing, Document Categorization, Clustering, Potential, Dendrite.

1 Introduction

In many cases, it is necessary to divide the set of the documents in a large document set or document flow into smaller groups and, instead of considering the whole group, to choose one representative in each such group, say, for a closer examination.

In different situations, different elements are to be chosen as representative in the group. Here are some practical examples:

- Case A: the "typical" (*average*) element is the most similar to all other elements.

  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. Using our program, one can 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 the 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 not only is to belong to his or her group but also should be most familiar with the other points of view. Using our approach, the organizer can cluster the proposals and assign the functions of the moderator of each group to the author of the document (proposal) with the desired properties automatically selected in the corresponding cluster.

- 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) of what Chinese people normally eat (because that would be rice) but the "least European" food from what Chinese people eat. Similarly, the "typical" (less Western) Russian wear is sarafan while "average" would be jeans. This kind of "the most typical" element is good to illustrate the diversity and emphasize the differences between groups. In a set of documents, one can be interested in reading

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

the ones that show the diversity of the clusters and do not contain much intersection.

Note that while in the example A the selected document is the nearest 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," as seen in Fig. 1.

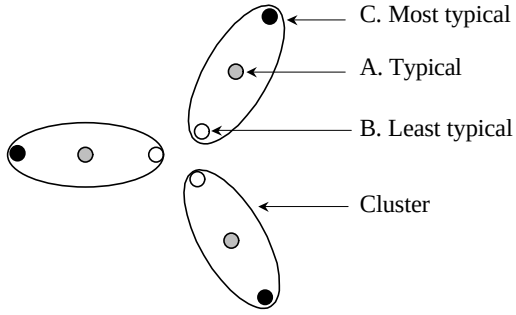

Fig.1 Types of representative elements

To compare documents by their closeness to a given domain or to evaluate the closeness between the documents, the following procedure is generally used: All documents are transformed to some numerical representation (called *document image*) and then the metric relations are introduced to obtain the quantitative estimations of the closeness.

Usually, the vector form of the document image is used, which is based on a given list of keywords. It allows to consider various measures such as polynomial, correlative, etc., and to use them in cluster analysis. Such analysis can be made, for example, by the method of the *nearest neighbor*, which is very popular in Text Mining. This 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 the document set reflecting a specific sub-domain. There is extensive literature discussing such methods; see, for example, the well-known monograph [6].

We also use such an approach. However, unlike typical applications, our program allows to form the combination of measures that reflects more exactly the closeness between two documents. Besides, we introduce new metric relations on dendrite based on the notion of *potential*. It allows taking into account not only the relation between the adjacent documents but between all ones.

The paper is organized as follows. In Section 2, we explain the kind of numerical representation we use for the text documents. In Section 3, we explain our main idea: the choice of the typical documents, which we then illustrate on an example in Section 4. Finally, in Section 5 we give the conclusions and discuss the possible future work.

2 Numerical Characteristics of the Documents

2.1 Document Image

We use the term *keyword* to refer to any key expression that can be a single word or a word combination. We represent a keyword in some normalized form reflecting a whole group of words with equivalent meaning. For example, a group of words *movement*, *movements*, *moving*, *move* are represented using a pattern `mov-`. For simplicity, hereafter we will still call such a pattern a keyword.

A domain dictionary (DD) is a dictionary consisting of such keywords (i.e., patterns) supplied with the coefficients A_k of importance for the given topic (domain). A coefficient of importance is a number between 0 and 1 that reflects the fuzzy nature of the relationship between the keywords and the selected domain. In other words, a DD is a fuzzy set of keywords.

In the simplest case, a user can build DDs using his or her own intuitive preferences. For user convenience, in our program the default value for such coefficients is 1. So, if a DD does not contain these coefficients, they all are considered to be 1.

However, statistical methods can be applied for automatically learning these coefficients. Since selection and classification results obtained using keyword lists are very sensitive to the contents of the DD, compilation of DD requires special technology and very careful work. An appropriate technology is described, e.g., in [5].

To apply a DD, for every document so-called *document image* is built. Consider the list of all keywords occurring in a given document. Let the number of occurrences of the keyword w_k in the document be n_k. Then the document image is formed by the values

$$X_k = A_k n_k,$$

where A_k, are the coefficients of importance for the corresponding keywords in the DD.

This set of values can be thought of as a vector in a multidimensional space. The direction of this vector (independently of its length) can be considered the *document theme*. This can be justified by the following consideration: a document consisting of several concatenated copies of a given document—which, obviously, reflects the same

theme—has the same direction of the document image vector, though a greater length.

2.2 Closeness between a Document and a Given Topic

Let (X_1, X_2, ..., X_N) be the image of a document for the given topic, where N is the total number of keywords in the DD. Then the *absolute document weight* can be calculated in different ways. The simplest variant is

$$W = \sum_{k=1}^{N} X_k \ , \qquad (1)$$

This measure has the important property of being additive with respect to the sub-domains reflected in the document. Namely, let a DD of N keywords is subdivided into two non-intersecting subDDs of N_1 and N_2 keywords, respectively, where $N_1 + N_2 = N$. As it was mentioned in the section 2.1 above, these subDDs define two different sub-topics. According to (1), the total amount of the keywords related to these sub-domains in the document equals to the total amount of the keywords related to the whole domain. This corresponds to the intuition about the contribution of sub-domains in their common domain.

In order to evaluate the correspondence of the document to the topic, the absolute document weight should be normalized by the document size. For this, if the document contains M running words (including keywords, but usually excluding stop-words like prepositions, etc.), then the *relative document weight* is W / M, where W is defined according to (1). This value is considered a measure of closeness between the document and a given topic.

2.3 Closeness between Documents

Let (X_{11}, X_{21}, ..., X_{k1}, ...) and (X_{12}, X_{22}, ..., X_{k2}, ...) be the images of two documents. If the two documents have the same length, then the following measures are used in literature to estimate the distance D between them:

- Correlative measure:

$$D = 1 - R, \qquad (2)$$

where R is the correlation coefficient:

$$R = \frac{\sum_k (X_{k1} \times X_{k2})}{\|X_1\| \times \|X_2\|}, \quad \|X_i\| = \sqrt{\sum_k X_{ki}^2} \ .$$

- Polinomial measure of various degrees:

$$D = \sqrt[p]{\sum_k (X_{k1} - X_{k2})^p} \ , \qquad (3)$$

where p = 1, 2, 4, ..., ∞.

The case p = ∞, obviously, corresponds to $D = \max_k |X_{k1} - X_{k2}|$. If the documents have different sizes M_1 and M_2, then the same formulas can be used, but instead of the coordinates X_{k1} and X_{k2} the following normalized values are to be used:

$$X'_{k1} = \frac{W}{M_1} X_{k1}, \ X'_{k2} = \frac{W}{M_2} X_{k2} \, . \qquad (4)$$

This normalization means that we reduce all estimations to the average of one word. Note that though the measures (3) and (4) have been extensively discussed in literature, in our case the coordinates of the vectors are scaled on the coefficients of importance, as explained at the end of the section 2.1.

The correlative measure is preferable if the user wants to evaluate the closeness between the themes of two documents. If the user wants to evaluate domain contribution in two documents then the polynomial measure is to be used. In the latter case, by increasing the degree p the user can emphasize the contribution of large differences in the numbers of occurrences of a small number of keywords.

However, in practice it is often desirable to combine both considerations, i.e., to take into account both the closeness of themes and the closeness of domain contribution of two documents. This can be achieved with a combination of the measures:

$$D = \alpha \times D_c + \beta \times D_p, \text{ where } \alpha + \beta = 1. \qquad (5)$$

Here, D_c and D_p are the distances between the documents in the correlative and the polinomial measures, correspondingly, defined according to (2) and (3), and α and β are the coefficients of preference—the penalties for the difference in the themes and in the relative domain contributions, correspondingly. Usually in practice we set α = β = 0.5. The combination of measures is discussed in [1].

Combining various measures, it is necessary to scale them to the same interval. For this, the

polynomial measure is corrected using the following scale coefficient:

$$C_S = 1 / \max D_{p.} \quad (6)$$

The maximum possible value of D_p is the maximal coefficient A_k: $\max D_p = \max_k A_k$. This equality takes into account the normalization (4). We assume that the combined measure (5) is applied to the normalized values.

3 Choice of Representative Documents

3.1 Clustering

Consider a non-oriented graph reflecting the relations between the documents. For example, this graph can be built by the method of the nearest neighbor. Note that the graph contains neither loops nor multiple arcs.

Definition 1. The *degree* of a node is the number of arcs connected with this node.

The nodes having degree 1 are called *hanging* nodes. The *length* of the arc is the distance between the two nodes it connects and can be calculated using the formulas from the section 2.3.

Let us now consider two non-adjacent nodes. By the distance between them, we mean the length of the shortest path. This path can be easily found on the dendrite because the dendrite is a tree (does not contain cycles).

Definition 2. The *potential* of a node p relative to some other node q is

$$Đ_{p,q} = \hat{e}_p / (1 + d_{p,q}), \quad (7)$$

where $\hat{e}_p$ is the degree of the node p and $d_{p,q}$ is the distance between the nodes p and q.

To avoid any influence of the hanging nodes, instead of (7) another formula can be used:

$$Đ_{p,q} = (\hat{e}_p - 1) / (1 + d_{p,q}). \quad (8)$$

For the graph presented in Fig. 2 assuming all arc length to be 1 and using the formula (7), we have $P_{e_1,a} = 1/5$, $P_{d,a} = 6/4$, $P_{c,a} = 2/3$.

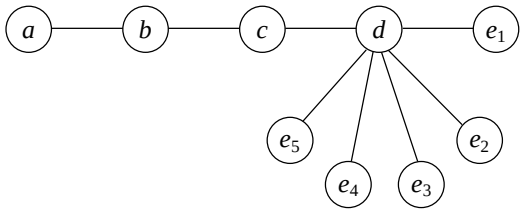

Fig.2 Example of graph

Definition. Potential of a node p is the sum of the potentials of all other nodes relative to the given one:

$$S_p = \sum_q P_{p,q} \ . \quad (9)$$

Definition. Difference of potentials between two adjacent nodes is the value defined as:

$$U_{p,q} = | S_p - S_q |. \quad (10)$$

For example, for the graph of Fig. 2, using the formula (8) we have S_a =25/12, S_b = 13/6, S_c = 3, so $U_{a,b}$ = 1/12 and $U_{b,c}$ = 5/6.

To build the clusters, the user should specify a threshold for the admissible level of the difference of potentials. Then the program eliminates all arcs that have a higher value of this difference, dividing one graph into several connected components. Each of these new graphs can be considered as a cluster.

High level of subjectivity in our definitions may set forth the problem of cluster validation. It may be particularly important to check the presence of the class structure. In this paper, we do not consider this question; for some suitable methods see [3, 4].

3.2 Choice of the Representative of a Cluster

After the document collection has been subdivided into clusters, it is possible to choose a representative 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; see the examples in the Introduction. As we have mentioned, for different tasks (see Introduction) different representatives are to be chosen. Accordingly, given a specific cluster, different criteria for the choice of the representative document can be suggested:

Case A. *Maximal closeness to the other documents in the cluster*.

This criterion implies calculating the potential of every document in the cluster; the document with the maximal potential is chosen. Here only documents in a given cluster are used in the sum in formula (9).

Case B. *Maximal closeness to the domain*.

This criterion implies calculating the relative document weight reflecting domain contribution to the document; the document with the maximum weight is chosen. Here the formula (1) is used with normalizing on the document size, as discussed at the end of Section 2.2.

Case C, variant 1. *Maximal distance from the domain.*

The calculation is as in Case B, but the minimum (instead of maximum) value is chosen.

Case C, variant 2. *Maximal distance from the documents in the other clusters*.

The calculation is as in Case A, but the minimum (instead of maximum) value is chosen, and the summation in the formula (9) is done by all documents of the collection except the ones pertaining to the cluster in question.

The difference between the results obtained for the cases A, B, and C is illustrated in Fig. 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 Fig. 3.

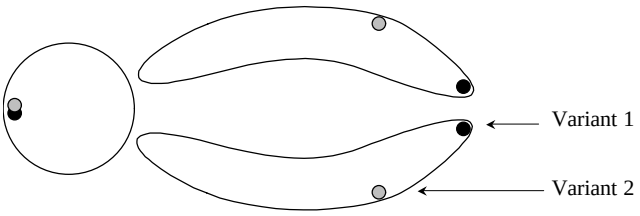

Fig.3 Types of representative elements

In our program, the user can select one of these criteria according to the specific task. For instance, the first criterion can be used in the situation described in the first example from Introduction—the choice of tests 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, if the organizer needs to select a representative from each group for negotiations between the groups for easier achieving consensus (the second example from Introduction), the persons reflecting the problem in the whole should be selected in each group. In this case, the second criterion can be used.

Very often in a cluster there are several documents that satisfy the selected criterion. In our implementation, in this case all such documents are checked using another criteria: for the case A, the criterion of the case B is used and vice versa, and for the case C, the other variant of the criterion is used. If after this several equal candidates still remain, then an arbitrary one is chosen.

4 A Practical Example

We have analyzed the papers on the medical domain kindly provided us by the specialists from the Masaryk University, Brno, Czech Republic. The test collection contains 711 papers. We have selected 124 papers for our experiments; the selection procedure has been described in [2]. We knew in advance that there were papers on cardiology, urology, therapeutics, etc., in total 8 sub-domains. Upon application of our program, we expected to obtain approximately 8 typical documents.

First, we constructed the DD for all document collection according the methodology described in [5]. Then, we clustered the documents trying various thresholds for the difference of potentials. Only using the correlative measure we could find some value of the threshold for which the number of clusters was close to 8 (namely 9); with the polynomial measures did not give the desired number of clusters. This means that the documents contained much noise information that was filtered out by the correlative measure. We chose the criterion of maximum closeness to the other elements of the cluster (case A) for the selection of typical documents. The results are presented in Fig. 4.

The scale (slider control) in the left-hand part of the program window allows the user to select the desired level of the threshold for admissible level of the difference of potentials.

It can be seen in the figure that the program discovered 9 non-elementary clusters. Their representatives are the files with the names 425-186.txt, 425-356.txt, 425-124.txt, etc.

The first cluster, which contains the greatest number of documents, was related to therapeutics.

Elementary clusters are those containing only one document. We obtained 10 such clusters. Their documents consider some specific issues so that they have very weak relation to other documents in the document collection.

The results of clustering coincided with the opinion of the human expert we consulted in 80% of the cases. But as for the representatives of the clusters, the documents selected by the program were within 3 best candidates selected by the human expert for a given cluster. However, almost in no case the automatically selected document coincided with the one ranked highest by the human expert. This problem is to be addressed in our future work.

5 Conclusions and Future Work

A domain dictionary consisting of domain-specific keywords gives a possibility to build various numerical measures for evaluation of closeness between a document and a given domain and between two documents. A combination of measures allows

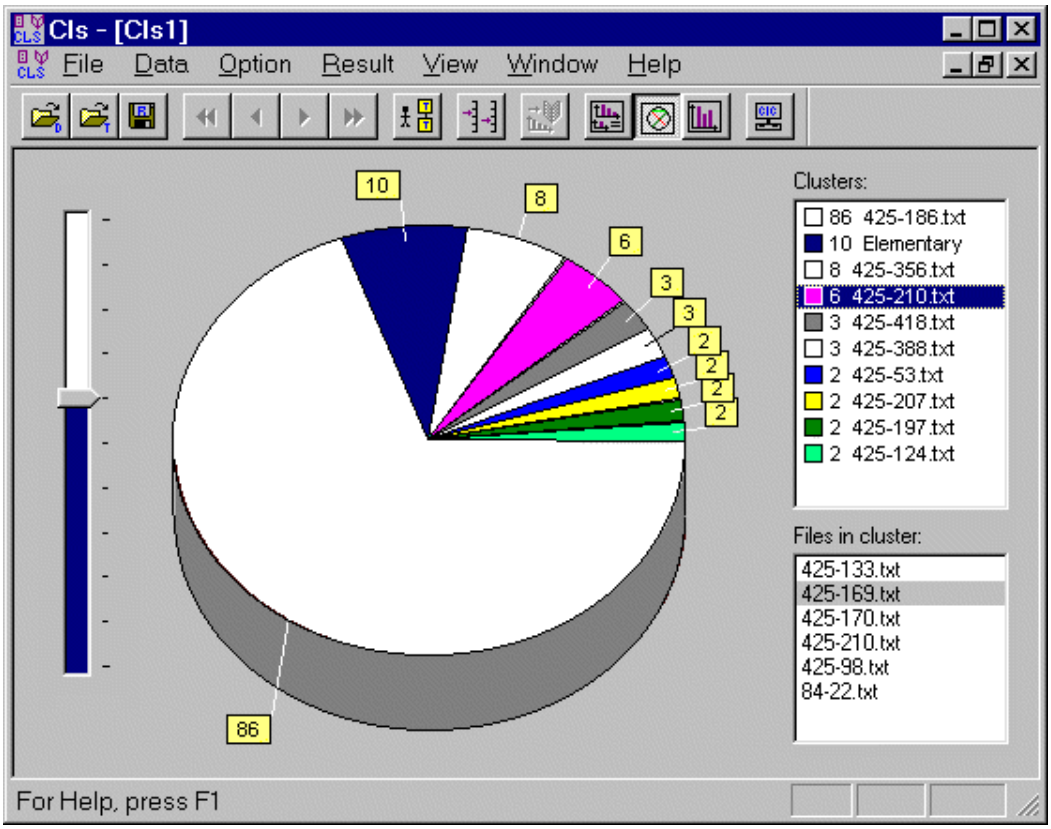

Fig.4 Clusters and their leaders in a medical document flow

taking into account the user preferences and the specific task the user is to accomplish.

For clustering on the graph of connections and for selection of the representative documents, the notion of potential has been proposed. It allows taking into account in more detail the graph structure.

The suggested techniques have been implemented in the program Document Classifier, which allows processing of large document collections.

In the future, we plan to increase the number of tuning parameters, for example, give the user the possibility to select an approximate desired number of clusters, degree of uniformity of their sizes, etc. Also, we plan to improve visual representation of the results.

References:

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

[2] M. Alexandrov, A. Gelbukh, and P. Makagonov. *Evaluation of Thematic Structure of Multidisciplinary Documents*. Proc. NLIS-2000, 2nd Int. Workshop on Natural Language and Information Systems at DEXA-2000, 11th Int. Conf. on Database and Expert Systems Applications, IEEE Computer Society Press, 2000. pp. 125–129.

[3] Gordon, A.D. *Cluster Validation*. Proceedings of the 5th Conf. of IFCS Data Science, Classification and Related Methods (Kobe, Japan), Springer-Verlag, 1996.

[4] Joliffe, I.T. *et al*. *Stability and influence in cluster analysis*. In: E. Diday (Ed.). Data Analysis and Informatics, v. 5, Amsterdam, 1988.

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

[6] Manning, D. C., H. Schutze. *Foundations of statistical natural language processing*. MIT Press, 1999.