Annales UMCS Informatica AI 2 (2004) 401-411

Annales UMCS
Informatica
Lublin-Polonia
Sectio AI

http://www.annales.umcs.lublin.pl/

Documents Clustering techniques[+]

Łukasz Machnik[*]

Department of Computer Science, Warsaw University of Technology,
Nowowiejska 15/19, 00-665 Warsaw, Poland

Abstract

Documents Clustering is a technique in which relationships between sets of documents are being automatically discovered and documents are divided into groups of similar specimens. The groups that are created during the process of clustering should be specified by the high degree of similarity between the elements that belong to the same group and low degree of similarity between the elements that belong to different groups. Such way of organizing documents allows the user to review content quickly and makes it easier to retrieve particularly interesting information. The following article describes the most popular documents clustering techniques and issues associated with it, like: text documents representation and similarity measure of documents. Additionally, the author is going to introduce his own concept of new effective method of documents clustering based on Ant System.

1. Introduction

Within the last few years the dynamic growth and expansion of Internet could be observed. Net is becoming a main publishing medium. The ability to access to such wide repository brings its users the invaluable benefits, however the problems with efficiency to retrieve information still occur.

Additionally, simple processing of natural language texts, based on retrieving and elimination of data by the comparison with a model, now is not enough for the users.

It is expected that data processing systems should support user's work by performing a full analysis of information meaning [1].

The described facts above were sufficient enough to increase people's interest in the natural language processing (NLS) area. It was expected that the progress in this area would help to improve the information retrieval in the Internet or would make it easier to arrange big data sets automatically.

[+] After second revision.

[*] *E-mail address*: L.Machnik@ii.pw.edu.pl

Another element that had a great influence on the increase of the interest in NLS was this new and popular trend related with efficiency in using knowledge in the organizations [2]. This trend was named: "knowledge management". It is generally known that text documents are the basic way of storing information in the organizations. That's why the new and optimal techniques of document management are still needed.

There are a few areas where methods of natural language processing are used:

- Analysis of structure, dependencies and rules which occur in the whole natural language (issues of language research, designing and building of language tools, like thesauruses, dictionaries and grammatical analyzers).
- Processing of single, text documents (automatic translation, key-word search, generation of abstracts).
- Analysis of text documents sets (search systems based on text data bases, documents clustering, classification and visualization).

2. Documents Clustering

Clustering analysis is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics and numerical analysis [3]. Van Rijsbergen crated the cluster hypothesis: "closely associated documents tend to be relevant to the same information requests" [4].

Documents clustering has been investigated for use in a number of different areas of text mining and information retrieval. Initially, document clustering was investigated for improving the precision in information retrieval systems. At present documents clustering is used for browsing a collection of documents and organizing the results, that were returned by a search engine in response to user's query.

Incorrectly, clustering is often mistaken with automatic classification. During the classification process the list of categories for which elements are assigned, is known before starting the processing. However, during clustering, the system does not have any initial knowledge and its task is to cluster elements into new categories. Elements that belong to the same category should be most alike, and in the same time they should differ meaningly from the elements in other categories [5]. This is the main difference between classification and clustering.

Clustering, contrary to the classification, is a process of machine learning without inspection.

3. Text document representations

Before describing the selected method a few essential topics should be discussed, for better understanding of documents clustering. At the beginning we should define a way of text document representation.

In the knowledge discovery area, the method of objects representation should base on the set of attributes that describe the sample. The text document is a series of words, and that is the reason why the transformation into the form that can be processed by the clustering algorithm is necessary.

For clustering algorithms, documents are represented using the vector-space model. In this model, each document, c, is considered to be a vector, $\boldsymbol{c}$, in the term - space (set of words which occur in all documents).

$$c = (\boldsymbol{t}_1, \boldsymbol{t}_2, ..., \boldsymbol{t}_n)$$

$\boldsymbol{t}_n$ – is the frequency of the n term in the document.

Normally very common words are stripped out and different forms of a word are reduced to one cononical form. It is the simplest and commonly used method of representation. It is called *uni-gram* representation.

There are two ways of creating a vector:

- Binary (only the fact of word appearance (or not) in the document is registered in the vector).
- Frequent (the frequency of the term appearance in the document is registered in the vector).

This method does not lack any disadvantages. Typical documents contain hundreds to thousands of different words, which causes many problems related to multidimensional space of attributes.

Therefore, researchers are still looking for a new solution to reduce this problem. Choosing only the most essential words or whole sequences is considered. A good example of such activity is γ – *gram* representation [6]. There are many methods that respect order, sequence of the words and their position in the document: *positional representation, n – gram representation*.

4. Similarity measure of documents

If a clustering algorithm is to be used the similarity between two documents must be measured in some way. It can be done as a measure of distance, as well as a measure of nearness (similarity).

As it was said earlier in this article, the most popular carrier of content similarity is frequency of terms in separate documents in the set that is being processed. The frequency of appearance is treated as the term weight in the document.

If we treat the set of these weights as **k** – the dimensional vector (**k** – the number of different words from all the documents in the examined set), we will obtain a vector model of a document with **k** attributes. Using these vectors and

possible measurements for computing the similarity between documents allows to calculate the content similarity $\mathbf{d(c_1,c_2)}$ between the documents $\mathbf{c_1}$ and $\mathbf{c_2}$.

These are the most popular measurements of similarity between documents:

– cosinus:

$$S = \cos(\mathbf{c}_1, \mathbf{c}_2) \approx \sum_{i=1}^{k} (\mathbf{p}_i \cdot \mathbf{q}_i),$$

– Euclidean distance:

$$S = \sqrt{\sum_{i=1}^{k} (\mathbf{p}_i - \mathbf{q}_i)^2},$$

– Square Euclidean distance:

$$S = \sum_{i=1}^{k} (\mathbf{p}_i - \mathbf{q}_i)^2,$$

– Manhattan:

$$S = \sum_{i=1}^{k} |\mathbf{p}_i - \mathbf{q}_i|,$$

– Chebychev:

$$S = \max |\mathbf{p}_i - \mathbf{q}_i|,$$

$\mathbf{p}_i$ – i-th element of $\mathbf{c}_1$ vector,
$\mathbf{q}_i$ – i-th element of $\mathbf{c}_2$ vector.

The presented measurements are usually used for calculating the value of the similarity coefficient based on the whole content of the document. Problems may appear in the Internet area, where only a part of document is often accessible.

5. Clustering techniques

Based on the structure of clustering results, clustering techniques can be easy classified as:

– Hierarchical,
– Non-hierarchical (partitional).

Hierarchical techniques produce a nested sequence of partitions, with a single, all-inclusive cluster at the top and single clusters of individual points at the bottom. Each intermediate level can be viewed as a combination of two clusters from the next lower level. The result can be displayed as a tree [7].

There are two basic approaches to generating a hierarchical clustering:

– Agglomerative start with the points as individual clusters and, at each step, merge the most similar pair of clusters until one cluster remains,
– Divisive start with one, all-inclusive cluster and in each step, split cluster until only single cluster of individual points remains.

In contrast with hierarchical techniques, partitional clustering methods create one – level (un – nested) of partitions. Basic non – hierarchical techniques divide input set into groups (covering of cluster is not allowed). However, there are a few methods that allow to assign a single element to more than one group.

6. Non-hierarchical clustering methods

The goal of non-hierarchical method activity is usually to find the best approximation. It can be achieved by dividing the set of documents into several initial clusters and then changing assign of elements until the expected results are achieved. Most of these methods are heuristic because the user has to define a number of result groups, criterion of document attachment and cluster representation at the beginning [8].

6.1. Single-pass methods

Algorithm:

1. Assign the first document D_1 as the representative of cluster C_1.
2. Calculate the similarity S_j between the document D_i and each cluster, keeping track of the largest, S_{max}.
3. If S_{max} is greater than $S_{threshold}$, add the document to the appropriate cluster, else create a new cluster with centroid D_i.
4. If documents remain, repeat starting from step 2.

Similarity: Computed between input and all representatives of existing clusters.

Terminology: Assumes *N* documents and *M* clusters.

Time: O(N log N)

Space: O(M)

Advantages: Simple, requiring only one pass through data; may be useful as a starting point for reallocation methods.

Disadvantages: Early in the process large clusters are produced; formed clusters are dependent on the order of input data.

6.2. Reallocation methods

Algorithm:

1. Select M cluster representatives or centroids.
2. Assign each document to the most similar centroid.
3. Recalculate the centroid for each cluster.
4. Repeat steps 2 and 3 until there is little change in cluster membership from pass to pass.

Similarity: Allows various definitions of similarity / cluster cohesion.

Terminology: Assumes *N* documents and *M* clusters.

Time: O(MN)

Space: O(M + N)

Advantages: Allows processing of larger data sets than in other methods.

Disadvantages: Can take a long time to converge into a solution, depending on the appropriateness of the reallocation criteria to the structure of the data.

The most popular representative of non-hierarchical techniques is *K – means* method.

7. Hierarchical clustering methods

In the early phase of research on documents clustering methods, the evolution of hierarchical techniques was limited by the calculation power of computers.

Together with the increase of the computers efficiency, the research on hierarchical methods gathered speed.

The goal of hierarchical algorithms is to produce a nested sequence of partitions.

As it was mentioned earlier, there are two types of hierarchical methods: agglomerative and divisive. Agglomerative techniques are more popular and gladly developed by the researchers.

The general agglomerative clustering algorithm is presented below:

1. Compute the similarity between every pair of accessible documents (similarity matrix).
2. Merge the most similar elements, replacing them with a single new point.
3. Update similarity matrix.
4. Repeat steps 2 and 3 until single cluster remains.

7.1. Single link method

Algorithm:

- Arrays record cluster pointer and distance information.
- Processing input documents/clusters one by one, for each:
  - Compute and store a row of the distance matrix.
  - Find the nearest other point using the matrix, and join this pair (single document or cluster) into a new cluster.
  - Relabel clusters, as needed.

Similarity: Joining the most similar pair of objects that are not yet in the same cluster. Distance between 2 clusters is the distance between the closest pair of points, each in one of the two clusters.

Terminology: Assumes *N* documents and *M* clusters.

Time: Usually O(N^2) though can range from O(N log N) to O(N^5).

Space: O(N).

Advantages: Theoretical properties, efficient implementations, widely used. No cluster centroid or representative is required, so no need arises to recalculate the similarity matrix.

Disadvantages: Unsuitable for isolating spherical or poorly separated clusters.

7.2. Complete link method

Algorithm:

- Arrays record cluster pointer and distance information.
- Processing input documents/clusters one by one, for each:
  - Compute and store a row of the distance matrix (descending order).
  - Find the nearest other point from the most distant one in the clusters, using the matrix, and join this pair (single document or cluster) into a new cluster.
  - Relabel clusters, as needed.

Similarity: Joining the least similar pair between each of two clusters.

Terminology: Assumes *N* documents and *M* clusters.

Time: Worst case is O(N^3).

Space: Worst case is O(N^2).

Advantages: All entries in a cluster are linked to one another within some minimum similarity.

Disadvantages: Difficult to apply to large data sets.

7.3. Group average link method

Algorithm:

- Processing input documents/clusters one by one, for each:
  - Find the nearest document/cluster (similarity between clusters means similarity between all the documents in the cluster) and join this pair into a new cluster.
  - Relabel clusters, as needed.

Similarity: Using the average value of the pair wise links within a cluster, based upon all objects in the cluster.

Terminology: Assumes *N* documents and *M* clusters.

Time: O(N^2).

Space: O(N).

Advantages: Ranked well in evaluation studies.

Disadvantages: Expensive for large collections.

Only the most popular hierarchical techniques were presented above. Other hierarchical techniques, like centroids or Ward's, are not described in this article [9].

8. Method of documents clustering based on Ant System

Generally the idea of Ant Systems was drawn from the observations of ants. Ants (*Linepithaema humile*) are the insects that live in the community called colony. The primary goal of ants is the survival of the whole colony. A single specimen is not essential, only bigger community may efficiently cooperate. Ants possess the ability of such efficient cooperation. It is based on work of many creatures that evaluate one solution as a colony of cooperative agents. Individuals do not communicate directly. Each ant creates its own solution that contributes to the whole colony solution [10].

The ability to find the shortest way between the source of the food and the ant-heel is a very important and interesting behavior of the ant colony. It has been observed that ants use the specific substance called pheromone to mark the route they have already gone along. When the first ant randomly chooses one route, it leaves the specific amount of pheromone, which gradually evaporates. Next ants which looking for the way, will, with great probability, choose the route where they feel more pheromone and after that they leave their own pheromone there. This process is autocatalic – the more ants go in a specific way, the more attractive it is for the others.

The analogy between finding the shortest way by ants and finding documents is similar (the shortest way between documents), and in addition, ability to use agents who construct their individual solutions as an element of the general solution, makes this idea very interesting and worth researching in the document clustering area.

An attempt to create a method of classifying text documents based on the artifical ant system has been considered. Application of such solution will be used as a method of finding the shortest path between the documents, which is the goal of the first phase (trial phase) of the method under consideration. The second phase (dividing phase) will have a task to separate actually a group of documents alike.

Finding the shortest path connecting every document in the set will be equivalent to building a graph whose nods would make up a set of analysed dosuments. Alike documents would be neighboring nods in the graph, considering that the rank of the individual nods will fulfil the condition of being smaller or equal to 2, which means that in the final solution one of the documents would be connected to only two others (most alike) – each document in the designed solution would appear only once. Obtaining such a solution would mean the end of the first phase, known as *preparing*. In the following stage of the process it is necessary to separate a group of documents like in a sequence obtained in the first phase. At the moment it is considered to apply different variants of the execution of the second phase of the process.

Just for the needs of building an effective method of classifying of the documents it is necessary to make a choice of possible modification and

adjusting of the concepts specific to real ants, so as they could be used effectively to solve problems connected with text mining.

– A colony of co-operating, individual specimen.

Artificial ants build a solution by moving along the graph of a problem, from one document to the other. During each iteration **m** number of ants constructs a solution in **n** number of steps, using a propabilistic law of making a decision. In practice, when visiting a specific document **i** ant chooses the next document **j** to move to, a pair (**i**, **j**) is added to the solution constructed at the moment. This step is repeated until the ant builds a complete solution for the specific iteration. Considering the fact that this version of algorithm is serial, after each ant finds a solution, a specific iteration process of leaving a certain amount of feromone associated with a pair of documents follows. After that the ant dies. Yet new ants appear in its place, whose goal is to find a solution in the following iteration, leave a feromone and die. The pattern repeats till the best result is reached or until a specific number of iterations is performed.

– A feromone trace and its force to influence.

Of the available variants of leaving feromone on the path, the author has chosen a partial variant. The ants leave a feromone in a specific amount which equals a quotient of a constatnt and a length of a found path. In addition, the decay of the feromone follows after constructing all partial solutions – the sum of distances between all the visited documents. The communication feromone path is being changed while finding a solution to a problem just to show the experience gained by ants while solving the problem.

– Finding the shortest path.

Coordinate description of the location of the specific document in space will be a vector representing the frequency of occurrence in the document. To describe the distance between the documents a simple measure in multidimensional space will be used – the Euclides distance. Finding the shortest path will be represented by finding such a sequence of passing from one document to the other, that the sum of the distances (measures of similarity) between the following elements of the examined set are smaller.

– Accidental movement of individual ants in the starting phase of finding the path

Keeping this condition is necessary because in the starting phase of algorithm action the ants are not able to use the experiance of their predecessors. The feromone trace between individual documents is equal to the selected constant value. Such a situation forces fully accidental choice of the documents in the starting phase of finding the path.

– Artificial ants live in the artificial, discreet world, can move only from one to the other specific position – states of the discreet world.

The set of states between which agents can move will be defined as a set of vectors representing the individual documents. As assumed earlier, each

document will be represented by a vector based on frequency of appearance of the specific word in the examined text.

– The amount of the feromone left by the artificial ant is connected with quality function of so far achieved solution

The amount of the feromone left by ants is proportional to the quality of the solution they found: the shorter is the distance between the documents the bigger is the amount of the feromone left on the pairs of the documents – the documents used to create the solution. The issue that still cannot be forgotten is requirement to evaporate the feromone. It is also necessary to exclude the stagnation phenomenon, that means: choosing the same route too early by all ants.

– The artificial ants are equipped with the memory of covered states, which is supposed to prevent the multiple location of one ant in the same position (it is necessary because there is a possibility/danger that ants could fall into cycles, which could make finishing building of the solution impossible).

9. Summary

Clustering of big document sets may be classified as a complicated computing problem. The researchers are still looking for new solutions in this field. They explore new spheres to find new ideas. For instance, they often look for inspiration in mother-nature. In the last few years many techniques based on nature were investigated: genetic algorithms, neurons networks, Tabu-search algorithms and the most interesting for the author: Ant Systems algorithms.

The techniques listed above usually use a group of cooperative and competitive agents to find a final solution of a problem. The author of this article studies using of Ant Systems based algorithm to evaluate a new efficient documents clustering method.

The preliminary experiments have shown that this idea could be useful. The research is in progress and that is why, the results, conclusions and ways of implementation have not yet been fully published.

References

[1] Pelc T., *Model reprezentacji semantyki zdań prostych język naturalnego dla potrzeb przetwarzania komputerowego*, Postscriptum, Kwartalnik Szkoły Języka i Kultury Polskiej, Uniwersytet Śląski w Katowicach, 27-29 (1998)-(1999), in Polish.

[2] Teece, D., *Research Directions For Knowledge Management*, California Management Review, September (1998).

[3] Berkhin P., *Survey Of Clustering Data Mining Techniques*, Technical Report, Accrue Software Inc, (2002).

[4] Rijsbergen van J., *Information retrieval*, Butterworths, London, (1979).

[5] Sholom W., White B., Apte C., *Lightweight Document Clustering* , IBM T.J. Watsan Research Center, (2000).

[6] Gawrysiak P., *Automatyczna kategoryzacja dokumentów*, Ph.D. Thesis, Warsaw University of Technology, (2001), in Polish.

[7] Steinbach M., Karypis G., Kumar V., *A Comparison of Document Clustering Techniques*, University of Minnesota, Technical report, #00-034, (2000).
[8] Kazienko P., *Grupowanie dokumentów hipertekstowych na podstawie drzewa maksymalnych przepływów*, Ph.D. Thesis, Wrocław University of Technology, (2000), in Polish.
[9] *CL Algorithm Details*, http://ei.cs.vt.edu/.
[10] Dorigo M., *The ant systems: optimization by colony of cooperating agents*, IEEE Transactions on Systems, Man, and Cybernetics-Part B, 26(1) 29.