1998 International Computer Symposium
Workshop on Artificial Intelligence
December 17-19, 1998, N.C.K.U., Tainan, Taiwan, R.O.C.

DOCUMENT CLUSTERING USING PROBABILISTIC NETWORKS

Shih-Yao Yang and Von-Wun Soo
Department of Computer Science
National Tsing-Hua University, HsinChu Taiwan, R.O.C.
Email: soo@cs.nthu.edu.tw

Abstract

We use Bayesian Belief Networks to perform inference on the similarity between a document and a cluster. We introduce the idea of "similar region" to construct the Bayesian Belief Networks and use a heuristic method to select initial seed objects. Our method allows a document to either belong to exactly one cluster or more than one cluster. Experiments are conducted to compare our method against K-means clustering method on a set of one-line Chinese news extracted from internet.

1. Introduction

1.1The Document Clustering Problem

The task of document clustering is to divide a collection of documents into several groups in each of which consists of a set of documents that are alike, while documents from different groups are not alike. Many document clustering methods often must determine the number of clusters and cluster's indices in advance. And they often require a document to be assigned to a unique cluster. But in reality some documents may relate to many different topics, they should appear in more than one cluster. We need a more flexible document clustering method.

1.2.Applying Bayesian Belief Networks to document clustering

Applying Bayesian Belief Networks is a method of probabilistic reasoning. Over the last few years, Bayesian Belief Networks have been applied to problems in medical diagnosis [1], information retrieval [3], automatic indexing [6], and so on. Turtle was the first person to apply the Bayesian belief networks to information retrieval [3]. It used the Bayesian Belief Networks to measure the similarity between the document and the information requested. We adapt this technique to document clustering.

2. Bayesian Belief Networks

A Bayesian Belief Network [4] is an acyclic directed graph in which the node represents a random variable. If there is an arc from variable B to variable A, B is said to be a parent of A and there is a conditional probability representing the degree of relation between variables A and B. For a variable x , let π (x) be the x's parent set in the belief network. We must give the probabilities of forming P(x| π (x)), the conditional probability of x given its parent set π (x). In a Bayesian belief network, there is an implicit assumption that each node is conditional independent of its non-descendents given its parents. Suppose there are n nodes in the belief networks. These nodes are ordered, x_1, x_2, x_3,…,x_n, and the parents of a node come before the node in the ordering. Then applying belief network's conditional independence assumption means that :

$$P(x_i|x_{i-1},x_{i-2},\ldots,x_1)=P(x_i|\pi(x_i))$$

By the chain rule we have

$$P(x_1,\ldots x_n)=\prod_{i=1}^{n}P(x_i|x_{i-1}\ldots x_1)=\prod_{1}^{n}P(x_i|\pi(x_i))$$

3. Inference network model

In this section, we introduce the inference network model used in information retrieval [3]. Inference network model uses the Bayesian theory to match the

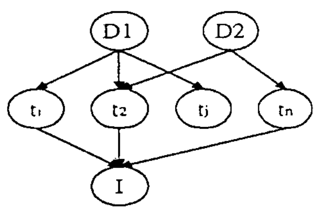

Figure 3-1

document and the information requested. Figure 3-1 is the architecture of the Bayesian information retrieval. It is a three-level hierarchy. The first level is the document in the collection, the second level is all the terms in the collection. If one term occurs in a document, then there is a direct link point form the document to the term. And on each link it exists a conditional probability representing the degree of relation between two nodes of the link. The third level is the information requested. If one term occurs in the information requested, there is also a link pointing from the term to information requested. After initialization the probability of requested information I given a collection documents D, I is computed as

$$P(I \mid D) = \sum_t P(I \mid t) \times P(t \mid D)$$

$$= \sum_t P(I \mid t_1, t_2, \ldots, t_n) \times \prod_{i=1}^{n} P(t_i \mid D)$$

The set of document retrieved given the information requested is the D that maximize P(I|D).

4. Bayesian Clustering

4.1 Introduction

Classifying objects according to similarities is the basis for Bayesian clustering. If each object belongs to exactly one cluster, we call it an exclusive classification, otherwise, nonexclusive or overlapping clustering. The Bayesian Clustering method we introduced can either be an exclusive or nonexclusive classification. The Bayesian Clustering is an unsupervised cluster. It automatically divides the collection of documents into several clusters, and the algorithm will select an initial seed object for each cluster. We introduce a method to automatically find the reasonable seed objects which is unlike the k-means method that randomly selects seed objects.

We also introduce a method of adjustment. In a collection of documents, some documents may not belong to any cluster at all. If we classify them into a cluster, they would become exception in the cluster. When the number of documents in a cluster reaches a threshold, we adjust the cluster by deleting the exception documents.

4.2 Determining the number of clusters

A significant problem in cluster analysis is the determination of the number of clusters in the final solution. Most clustering methods are not designed to determine the number of clusters in the data. For K-means method, the user must specify a number of clusters first.

4.2.1.Similar region

We represent each document as a point in a graph. The distance between a pair of points represents the similarity between them. The shorter the distance, the greater is the similarity. Each point P in the graph has a similar region in which p's most similar points within a predefined similarity threshold are all in the region as Figure 4-1. Now we show how to find out each point's similar region.

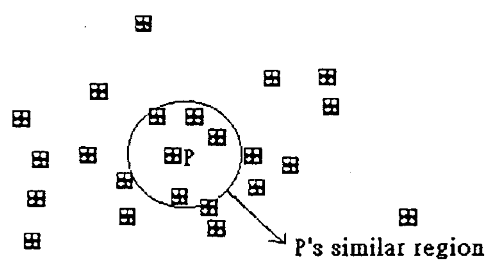

Figure 4-1

We use the vector space model (Salton 1988) to measure the similarity between two documents. We use the model to compute similarities between P and all other n points, and let them be S_1, S_2, ..., S_n. Then to estimate the standard deviation S and average A of S_1, S_2, ..., S_n.

$$A = \frac{S_1 + S_2 + \ldots + S_n}{n} = \frac{\sum S_i}{n}$$

$$S = \frac{\sum_{i}^{n}(S_i - A)^2}{n}$$

For each point P_i, S_i is the similarity between P and P_i, if $S_i > \alpha * S + A$, then P_i is within the P's similar region. The value of α is a parameter to control the size of the similar region.

4.2.2.The algorithm to determine the number of clusters

In this section, we introduce an algorithm to determine the number of clusters in the final solution.

Definition (similarity frequency): For each point P, P's similarity frequency (SF_p) is the number of similar regions that contain point P.

The algorithm first computes SF for all points. The bigger the SF of a point p suggests a higher tendency that p may be the center point of some set of points in the graph. So the algorithm first chooses point i with the maximal SF as the seed point of the first cluster. Then it chooses point j with the second maximal SF that isn't within the similar region of i as the second seed point of the second cluster. Repeat the process until no points are left out. Finally, the algorithm will generate a set of clusters which each has one seed point.

We assume the algorithm can generate n clusters at most. If human initialially gives a number m (m>n), then it will just generate n clusters in the final solution. On the contrary, if m<n, then it will generate m cluster.

4.3.Architecture

In the above section (4.2.2), we introduce an algorithm to roughly generate n clusters in which each has a seed point. In this section, we introduce an architecture that measures the similarities between object and each cluster and classifies the objects into appropriate clusters.

4.3.1The Network Model

Figure 4-2 shows the architecture of a Bayesian Belief Networks with two clusters C1 and C2. It is a three-level hierarchical network. The first and second levels are the

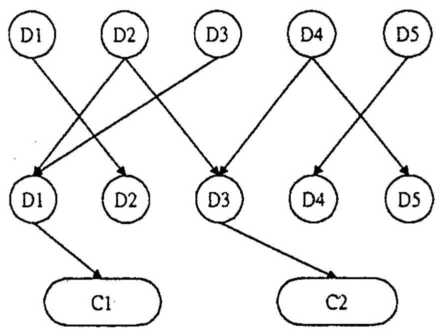

Figure 4-2

documents in the collection. Each document in the first level has some links direct to the documents in the second level. In section 4.2.1, we introduce a method to find out the similar region for each document. If we have found out the similar region of document D, then there are links pointing from D to the documents within D's similar region. In Figure 4-2, we assume the system generate two clusters C1 and C2 initially with seed points D1 and D3 respectively. Suppose D1's similar region contains D2, D2's similar region contains D1 and D3, D3's similar region contains D1, D4's similar region contains D3 and D5, and D5's similar region contains D4.

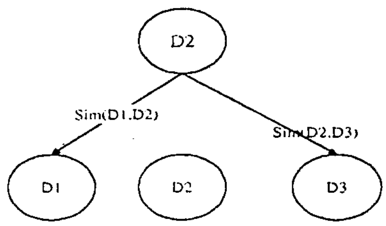

Figure 4-3

On each link we attach a conditional probability to represent the degree of similarity as shown in figure 4-3. In our system, we calculate the conditional probability as follows:

$$P(D1 \mid D2) = Sim(D1, D2)$$

where Sim(D1,D2) is calculated by vector space model (Salton 1988).

If document D belongs to Cluster C, then there is a link pointing D to C. If D belongs to many clusters, then there is a link from D to each cluster. If cluster C contains n documents, then there are n links pointing to cluster C.

4.3.2.The linking probability of document nodes to a cluster node

In Figure 4-4, the cluster node C1 contains three documents D1, D2, D4. We assume that the belief

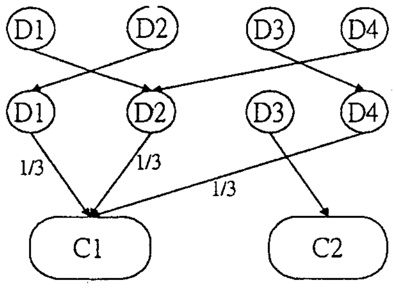

Figure 4-4

of C1 depends on the parents that are true and all parents are weighted equally, so the conditional probability P(C1|D1), P(C1|D2), P(C1|D4) are all set as 1/3.

If the probability of D1, D2, and D4 are P_1, P_2, and P_4, then the probability of C1 is computed as follows: P(C1=true) = (P1+P2+P4)/3.

4.3.3. Probabilistic inference

Given a document D, how to estimate the probability of D with each cluster. In Figure 4-5, cluster C1 contains D2, and D4, and cluster C3 contains D5, D6, and D7. Now we will classify document D3 into an appropriate cluster. Given the initial node D3, we can calculate the conditional probabilities of cluster C1 and C2 respectively.

$$P(C2 \mid D3) = \frac{0.8+0+0.7}{3} = 0.5$$

$$P(C1 \mid D3) = \frac{0.7+0}{2} = 0.35$$

According to the probabilities, classifying document D3 into the closest cluster if we allow a document to exactly belong to one cluster. In Figure 4-5, we classify D3 into cluster C1. In non-exclusive clustering, we can

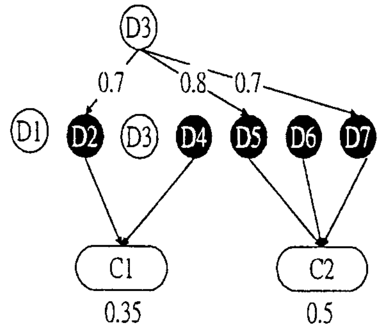

Figure 4-5

set a threshold and allow any document to put in a cluster whose similarity to the cluster is above the threshold.

4.3.4.Classifying a document into a cluster

Given a document, determine which cluster the document belongs is the classification. Assume there are n clusters the system generated, and the similarities between a document and n clusters are S_1, S_2, ..., S_n respectively. In exclusive clustering, we just add the document to the cluster whose similarity is the greatest. However, in non-exclusive clustering, a threshold must be properly chosen. If S_i > threshold, the document is added to cluster i.

To determine the threshold, we first estimate the standard deviation S and the average A of S_1, S_2, ...S_n. Set the threshold θ as S + A. For all i if $S_i > \theta$, add the document to cluster i. If the document allow to belong to many clusters, the document may be in the boundary.

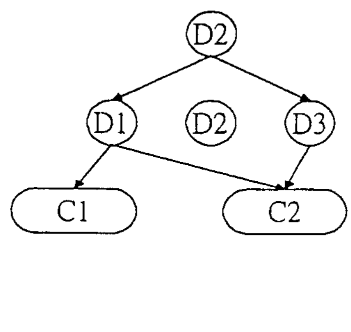

Figure 4-6

In Figure 4-6, D1 belongs to two cluster C1 and C2.

Now we will calculate conditional probabilities of C1 and C2 given document D2. Because D1 is within D2's similar region, so there exists a link between them.

In section 4.3.1, we set the conditional probability as follows:

$$P(D1 \mid D2) = Sim(D1, D2)$$

Because D1 belongs to both cluster C1 and C2, so it inferences D2 loosely. We modify the conditional probability of D1 given D2 as follows:

$$P(D1 \mid D2) = \frac{Sim(D1, D2)}{\log(2+1)}$$

The term log(2+1) means that document D1 belongs to two clusters. The more clusters to which document D1 belongs, the less important is D1 in each cluster.

The general form is

$$P(D1 \mid D2) = \frac{Sim(D1, D2)}{\log(n+1)}$$

, where D1 is simultaneously contained in n clusters.

4.3.5.The Bayesian clustering algorithm

In this section, we describe a document clustering algorithm Bayesian Clustering in Figure 4-8 using Bayesian Belief Networks. It contains three input parameters, the *array of object*, the *number of cluster* in the final solution, and a boolean variable *Overlapping* to determine whether an object belongs to exactly one cluster or can belong to more than one cluster. If the parameter Overlapping is false, each object exactly belongs to one cluster, otherwise each object can belong to more than one cluster.

4.3.6.Adjustment

In Step 4 of the Bayesian Clustering algorithm, the algorithm randomly select a document and add it to a cluster one by one. For different order of selections, the same document may be classified into different clusters.

In figure 4-7, the algorithm first generates two clusters C1 and C2 that have seed points p and q respectively. Then the algorithm randomly selects a point s, because s is close to q so it will be added to cluster C1. But the point s may belong to cluster C2 if all other data points in C2's similar region are added first. So the algorithm must adjust this error. Point s must be deleted from cluster C1 and be added to cluster C2.

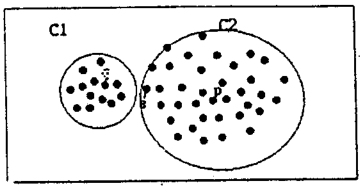

Figure 4-7 Point s will be added to cluster C1

Define (The center document of a cluster): Assume cluster C contains n documents, D1, D2, ...,Dn. For each Di, we compute the similarity between Di and cluster C. Let them be S_1, S_2, ...S_n. If S_c is the greatest among these values, then we let D_c be the center document of cluster C. Because the data may be removed from a cluster, so the program must traverse the data set and classify the unclassified data into an appropriate cluster again. "When to adjust a cluster C ?". When a document be assigned to cluster C, we calculate the center document D_c of the cluster C. Assume D_c's similarity frequency (SF) is m, we set a threshold θ equal to β*m. If cluster C contains n elements and n > θ, the algorithm will remove some elements that are not similar in cluster C. The similarity frequency (*SF*) m of D_c is the number of similar regions contains point D_c. We set β equal 1+ loop* δ where loop is the time the program traverse the data set and δ is positive. Because the δ is positive, so the threshold θ increases dynamically. We increase the threshold θ, because if the value is fixed, the program will adjust the cluster very frequently and some documents may be removed or added frequently. If the value of δ is large, the threshold will increase fast.

"What document will be removed from cluster C ?". If cluster C contains n documents, we first evaluate the similarity between each document and cluster C and get the similarities S_1, S_2, ..., S_m, and then evaluate the average A and standard deviation S of S_1, S_2, ..., S_m. We declare the remove threshold ε equal to α *S + A . For each S_i, if $S_i < \varepsilon$, then remove Di from cluster C, where α equal to 1-loop* γ and γ is positive, so both the value of α and the remove threshold ε decrease dynamically. Because if the value is fixed, the when a cluster C growth to a great size, there is not any document can be added to the cluster. If the value of γ is large, then the threshold will decrease fast.

If a document is removed from a cluster, and the document belongs to this cluster only, then we must mark the document as unclassified. Because documents can be removed from a cluster, some documents may not be assigned to a cluster while the program executes for some times. So we will jump the while loop when the program executes for a long time. The detailed algorithm of adjustment and the final algorithm of step 4 with adjustment are not shown due to the space limitation.

5.The Experiments

5.1.The Data Set

In this section, we describe the data set used. We collect 1584 documents from the WWW (World Wide Web). These documents are all news. It was originally taken from eight different classes. Table-1 shows the detail of the data set.

5.2.Preprocessing

Because the data set consists of all text documents, so we must perform IR's pre-processing. We must find out all the characters in the data set. Because all the documents are Chinese news, so we just process the Chinese character. We find out 4912 Chinese characters from the 1584 documents. Then compute the idf (inverse document frequency) for each character. We erase some character whose idf is very small and erase those that exactly present in one document. The final number of words is 2721. After finding out all different characters, we must evaluate the weight of each character in a document. Then use the vector space model to measure the similarity between each pair of document.

5.3Results

5.3.1.Experiment1

Table-1 The data set contains eight classes and has 1584 documents in total.

Class	大陸新聞	臺灣新聞	社會新聞	國際新聞	焦點新聞	經濟新聞	影藝新聞	體育新聞
Number	129	235	183	154	207	234	214	228

In this experiment, we tentatively classify the data set into eight clusters and each document exactly belongs to one cluster. Table-2 and Table-3 shows the result of the Bayesian clustering and K-means. In this experiment, Bayesian Clustering and K-means use the same seed documents. The seed document is generated by the procedure described in section 4.2.2. From the result of the Bayesian Clustering, the classes of "經濟", "影藝", and "體育" are almost found out. But the result of the K-means clustering found out two classes "經濟", and "體育" and it has one cluster is larger.

5.3.2. Experiment2

Table-2 The result of Bayesian Clustering. The seed documents are generated by the program

Cluster	大陸	臺灣	社會	國際	焦點	經濟	影藝	體育	Total
1	1	2	4	11	2	0	1	207	228
2	0	12	15	9	8	4	194	5	247
3	8	10	6	6	28	211	3	0	272
4	6	15	114	13	39	5	3	1	196
5	3	48	22	12	15	2	7	2	111
6	24	105	5	42	67	4	4	10	261
7	83	28	9	52	26	6	2	2	208
8	4	15	8	9	22	2	0	1	61
Total	129	235	183	154	207	234	214	228	1584

In this experiment, we also classify the data set into eight clusters and each document exactly be longs to one cluster. Table-4 and Table-5 shows the result of the Bayesian clustering and K-means. In this experiment, Bayesian clustering and K-means use the same seed documents. But the seed documents aren't generated by program. We selected eight seed documents from each class.

In this result, we use different seed documents. In the two

BayesianCluster(Array **object[]**, int **ClusterNumber**, boolean **Overlapping**)

Input : 1. objects consists of a set of documents,D1,D2,...,Dn.

2. The parameter **ClusterNumber** n as the number of cluster in final solution.

3. The parameter **Overlapping** is a boolean variable.

Output : A set of clusters.

Step 1: Evaluate all document's similar regions. (4.2.1)

Step 2: Evaluate all document's similar frequencies. (4.2.2)

Step 3: Determine the number of clusters and find a seed document for each cluster. (4.2.2)

Step 4: randomly select a document and add it to a cluster one by one

Figure 4-8 Algorithm of Bayesian Clustering

Table-3 The result of K-means . The seed documents are generated by the program.

Cluster	大陸	臺灣	社會	國際	焦點	經濟	影藝	體育	Total
1	0	0	0	2	0	0	0	0	2
2	9	11	7	1	34	209	4	1	276
3	0	2	1	0	0	0	0	0	3
4	1	9	11	11	6	1	84	201	324
5	3	2	2	22	2	2	0	7	40
6	5	1	2	0	1	1	0	0	10
7	1	1	3	0	1	0	3	3	12
8	110	209	157	118	163	21	123	16	917
Total	129	235	183	154	207	234	214	228	1584

Table-4 The result of Bayesian Clustering. The seed documents are selected from each class.

Clus-ter	大陸	臺灣	社會	國際	焦點	經濟	影藝	體育	Total
1	1	2	6	9	4	0	2	201	225
2	0	11	18	13	8	0	185	4	239
3	6	9	8	7	29	201	5	0	265
4	9	9	99	13	33	5	5	1	174
5	40	115	15	28	62	2	8	14	284
6	58	27	8	46	28	8	4	3	182
7	4	46	20	16	30	3	4	1	124
8	11	16	9	22	13	15	1	4	91
Total	129	235	183	154	207	234	214	228	1584

experiments, the results of Bayesian clustering have less variation than K-means clustering. The Bayesian clustering also found out three classes. The K-means clustering didn't generate large cluster. We observe that the performance of the K-means clustering depends on the initial positions of the seed documents.

Table-5 The result of K-means . The seed documents are all selected from each class.

Cluster	大陸	臺灣	社會	國際	焦點	經濟	影藝	體育	Total
1	3	33	15	3	35	185	2	2	278
2	3	2	1	21	1	3	0	1	32
3	9	21	15	81	18	9	67	207	427
4	4	1	2	0	1	1	0	0	9
5	17	81	5	9	73	1	5	8	199
6	31	14	5	1	15	2	7	2	77
7	5	35	110	14	37	5	119	5	330
8	57	48	30	25	27	28	14	3	232
Total	129	235	183	154	207	234	214	228	1584

6. Discussion

We implement document clustering using the Bayesian Belief Networks. Three parameters to be assigned, "when to adjust a cluster", "remove threshold" and "thesize of similar regions".

- **When to adjust a cluster?**

Because the algorithm randomly select a document to assign to a cluster. Different selection order may classify some documents into different clusters. So when a cluster reaches a moderate size, we remove some document from the cluster. How to decide the size is a problem. We assign the size a small value initially, then dynamically increase it. If we assign a big value initially, it will make some clusters grow fast and clusters will contain more dissimilar documents. But if we assign a small value initially, the size of a cluster will not grow fast and the intra-cluster similarity will be small.

- **Remove-threshold**

When we adjust a cluster, we must remove dissimilar documents from the cluster. But how to decide a document is dissimilar? We assign the remove-threshold a big value initially, then dynamically decrease it. If the remove-threshold is big, the inter-cluster similarity will be larger and it causes the size of cluster smaller.

- **The size of similar region**

The size of similar region determines the number of clusters to be generated. It affects the similarity frequency of document. The smaller the size, the greater the number of clusters.

7. Conclusion and Future work

7.1.Conclusion

In this paper, we introduce an architecture to perform document clustering. The architecture is a three-level hierarchy Bayesian Belief Network. We introduce two new ideas "similar region" and "similarity frequency" and use the former to construct the belief network and the later to select the seed documents and to determine the number of clusters in the final solution. In the experimental results, we used different seed documents for two experiments. We found the results of Bayesian Clustering were not significantly affected by the selection of initial seed documents while those of the K-means Clustering method were depending on the initial seed documents.

7.2.Future Work

In this paper, although the data set are divided into eight classes, many classes are similar. It is difficult to extract each class from the data set. So the selection of Corpus is very important. In the future, we can use this architecture to cluster other data set, such as picture or other of documents.

The Bayesian Clustering method is an unsupervised method, so it is difficult to assess the results. In the future, we can change this algorithm to supervised method and use it to do classification.

The Bayesian Clustering is either an exclusive or nonexclusive classification. But in this paper, we only show experimental results of the Bayesian Clustering as exclusive classification. This is because data set of nonexclusive classification is not available.

The current architecture only has three-level hierarchy, we can extend it to more levels or add more extra evidence or constraints.

Reference

[1] Heckerman (1990). Probabilistic Similarity Networks. Technical Report STAN-CS-1316, 1990.

[2] G. Salton and C. Buckley, Term weighting approaches in automatic text retrieval, Information Processing and Management 24(3), pp.513-524, 1988.

[3] H. Turtle and B. Croft, Inference Network for Document Retrieval, ACM_SIGIR, pp.1-24, 1990.

[4] J. Pearl , Probabilistic Reasoning In Intelligent Systems, San Mateo, CA: Morgan Kaufmann Publishers,1988

[5] R.R. Yager and D.O. Filev, Approximate clustering via the mountain method. IEEE Transactions on Systems, Man, and Cybernetics 24, pp.209-219, 1994.

[6] T. Kostas, H. Stephan Automatic Indexing Based on Bayesian Inference Networks, ACM-SIGIR , pp.24-34, 1993.