CLUSTERING OBJECTS ON SUBSETS OF ATTRIBUTES BASED TEXT CLASSIFICATION

S.Mohanraja[#1], V.Mohana Priya[*2], P.Savitha[*3], N.Vijayakumar[#4], Mrs.V.Sharmila[*5] (Assistant Professor)

[#]*Department of Computer Science and Engineering, Anna University*
K.S.R.College of Engineering, Tiruchengode-637 215,
Namakkal district, Tamil Nadu,India
[1] s.mohanraja93@gmail.com
[3]savitha6302@gmail.com[3]

[*] *K.S.R.College of Engineering, Tiruchengode-637 215,*
Namakkal district, Tamil Nadu,India
[2]priya040493@gmail.com

[4] vijayakumarcse793@gmail.com

ABSTRACT- In the text processing field, the important aspect is measuring the similarity between the documents. Initially by giving the keyword as input and it will collect all the relevant web pages. By applying the pre-processing method, it will remove all the stop words within the documents. Then the nodes and edges are created for each and every document using the parsing method. The similarity values will be calculated based on the TF-IDF measure which depends on the OLP values. The multi view point technique is used to cumulate the documents and form the clusters using clustering algorithm. If there is any additional documents, those documents will be either added to an existing clusters or forms a new cluster using incremental clustering method. Using multi view points, more informative assessment similarity could be achieved.

Keywords- Text processing, TF-IDF measure, OLP values, Multi view point technique, Parsing method, Incremental clustering.

INTRODUCTION

In information retrieval, data mining and web search, text processing plays an important role. The bag-of-words model is commonly used in text processing. The document is usually represented as a vector, in which each component indicates the value of the corresponding feature in the document. The feature Value can be term frequency can be the number of occurrences of all the terms in the document set, or tf-idf is a combination of term frequency and inverse document frequency. Most of the feature value in the vector are zero. The size of the dimensionality of the document is large and the resulting vector is sparse. Measuring similarity in high dimensionality and sparse are a very challenging task which is important in text processing algorithms. For computing the similarity between two vectors has a lot of measures. The difference between the probability distribution associated with the two vectors is a Kullback-Leibler divergence. The hamming distance between two vectors is the number of positions at which the corresponding symbols are different. The sparsity property of the cosine similarity measure is used to retain the extended Jaccard coefficient and the Dice coefficient. In text classification and clustering algorithms similarity measures have been extensively used. In the existing system, Clustering is one of the most interesting and important topics in data mining. The aim of clustering is to find intrinsic structures in data, and organize them into meaningful subgroups for further study and analysis. There have been many clustering algorithms published every year. It greedily picks the next frequent item set which represent the next cluster to minimize the overlapping between the documents that contain both the item set and some remaining item sets. In other words, the clustering result depends on the order of picking up the item sets, which in turns depends on the greedy heuristic. This method does not follow a sequential order of selecting clusters. Instead, we assign documents to the best cluster. This method does not follow a sequential order of selecting clusters. It provides minimum efficiency and performance. In the proposed system, the main work is to develop a novel hierarchal algorithm for document clustering which provides maximum efficiency and performance. It is particularly focused in studying and making use of cluster overlapping phenomenon to design cluster merging criteria. Proposing a new way to compute the

overlap rate in order to improve time efficiency and "the veracity" is mainly concentrated. Based on the Hierarchical Clustering Method, the usage of Expectation-Maximization (EM) algorithm in the Gaussian Mixture Model to count the parameters and make the two sub-clusters combined when their overlap is the largest is narrated. Experiments in both public data and document clustering data show that this approach can improve the efficiency of clustering and save computing time. Multi-view point, Reduces irrelevant data and it improves efficiency and performance.

USING SINGLE VIEW POINT

Some measures which have been popularly used for calculating the similarity between the two documents are briefly discussed here. Let $\mathbf{d}_1$ and $\mathbf{d}_2$ be two documents represented as vectors. The Euclidean distance [45] measure is defined as the root of square differences between the respective coordinates of $\mathbf{d}_1$ and $\mathbf{d}_2$, i.e.,

$$d_{Euc}(\mathbf{d}_1,\mathbf{d}_2)=[(\mathbf{d}_1-\mathbf{d}_2)\cdot(\mathbf{d}_1-\mathbf{d}_2)]^{1/2} \quad (1)$$

where $\mathbf{A}\cdot\mathbf{B}$ denotes the inner product of the two vectors **A** and **B**. Cosine similarity [25] measures the cosine of the angle between $\mathbf{d}_1$ and $\mathbf{d}_2$ as follows:

$$S_{Cos}(\mathbf{d}_1,\mathbf{d}_2)=(\mathbf{d}_1\cdot\mathbf{d}_1)^{1/2}(\mathbf{d}_2\cdot\mathbf{d}_2)^{1/2}. \quad (2)$$

Pairwise-adaptive similarity [17] randomly selects a number of features out of $\mathbf{d}_1$ and $\mathbf{d}_2$ and is defined to be,

$$dp_{air}(\mathbf{d}_1,\mathbf{d}_2)=d1.d2/(\mathbf{d}_{1,K}.d1,K)^{1/2}(\mathbf{d}_{2,K}.d\,K)^{1/2} \quad (3)$$

where $\mathbf{d}_{i,K}$ is a subset of $\mathbf{d}_i$, i = 1, 2, containing the values of the features which are the union of the K largest features appearing in $\mathbf{d}_1$ and $\mathbf{d}_2$, respectively.

The Extended Jaccard coefficient [48], [49] is an extended version of the Jaccard coefficient [21] for data processing:

$$S_{EJ}(\mathbf{d}_1,\mathbf{d}_2)=d1.d2/(d1.d1+d2.d2-\ d1.d2) \quad (4)$$

while the Dice coefficient looks similar to it and is defined as follows:

IT-Sim, an information-theoretic measure for document simi- larity, was proposed in [39], [8]:

$$S_{IT}(\mathbf{d}_1,\mathbf{d}_2)= \frac{2\sum_{wi} \min(p_{1i},p_{2i})}{\log\pi(w_i)}$$

$$\sum_{wi} p_{1i}\log\pi(w_i)+\sum_{wi} p_{2i}\log\pi(w_i)$$

$$S_{Dic}(\mathbf{d}_1,\mathbf{d}_2)=2d1.d2/(d1.d1+d2.\ d2) \quad (5)$$

where w_i represents feature i, p_{ji} indicates the normalized value of w_i in document $\mathbf{d}_j$ for j= 1 or j=2, and $\pi(w_i)$ is the proportion of documents in which w_i occurs.

USING MULTI VIEW POINT

By the usage of Multi view point technique, histogram, similarity and frequency are to be considered for the given input keyword. Those are checked within the related documents of the given keyword. Let a document **d** with m features $w_1, w_2, \ldots, w_m$ be represented as an m dimensional vector, i.e.,

$\mathbf{d} =< d_1, d_2, \ldots, d_m >$.

If w_i, $1 \le i \le m$, is absent in the document, then $d_i = 0$.

Otherwise, $d_i > 0$. The following properties, among other ones, are selected 1) The presence or absence of a correspondence value is more essential than the difference between the two values associated with a present document. Consider two correspondence value w_i and w_j and two documents $\mathbf{d}_1$ and $\mathbf{d}_2$.If w_i does not present in d_1 but it present in $\mathbf{d}_2$,then w_i is considered to have no correlation with $\mathbf{d}_1$ while it has some correlation with $\mathbf{d}_2$. In this case, $\mathbf{d}_1$ and $\mathbf{d}_2$ are not related in terms of w_i. If w_j appears in both $\mathbf{d}_1$ and $\mathbf{d}_2$. Then w_j has some correlation with $\mathbf{d}_1$ and $\mathbf{d}_2$ respectively. In this case, $\mathbf{d}_1$ and $\mathbf{d}_2$ are related to some level in terms of w_j.For the above two cases, it is practical to say that w_i carries an additional weight than w_j in determining the correspondence level between $\mathbf{d}_1$ and $\mathbf{d}_2$. For example, take for granted that w_i is absent in $\mathbf{d}_1$, i.e., $d_{1i} = 0$, but appears in $d_{2,}$ e.g., $d_{2j} = 2$, and w_j appears both in d_1 and $d_{2,}$ e.g., $d_{1j} = 3$ and $d_{2j} = 5$. Then w_i is considered to be more fundamental than w_j in determining the relationship between d_1 and d_2.

2) The correspondence level should increase when the difference between two non-zero values of a specific aspect decreases. For example, the similarity involved with $d_{13} = 2$ and $d_{23} = 20$ should be smaller than that in for a correspondence measure between two documents complicated with $d_{13} = 2$ and $d_{23} = 3$.

3) The similarity level should decrease when the number of presence-absence features increases. For a presence- absence feature of $\mathbf{d}_1$ and $\mathbf{d}_2$, $\mathbf{d}_1$ and $\mathbf{d}_2$ are not related in terms of this aspect

as demonstrated earlier. Therefore, as the number of presence-absence features increases, the dissimilarity between $\mathbf{d}1$ and $\mathbf{d}2$ increases and thus the similarity decreases. For example, the similarity between the documents $< 1, 0, 1 >$ and $< 1, 1, 0 >$ should be smaller than that between the documents <1,0,1> and <1,0,0>.

4) Two documents are less related to one another if none of the aspects have non-zero values in both documents. Let $\mathbf{d}1 = < d11 , d12 , \ldots , d1m >$ and $\mathbf{d}2 = < d21 , d22 , \ldots , d2m >$. If,

$$d_{1i} d_{2i} = 0,$$
$$d_{1i} + d_{2i} > 0$$

for $1 \leq i \leq m$, then $\mathbf{d}1$ and $\mathbf{d}2$ are least similar to each other. As mentioned earlier, $\mathbf{d}1$ and $\mathbf{d}2$ are not related in terms of a presence-absence feature. Since all the features are presence-absence features, the dissimilarity reaches the limit in this case. For example, the two documents $< x, 0, y >$ and $< 0, z, 0 >$, with x, y, and z being non-zero numbers, are least similar to each other.

5) The similarity measure should be symmetric. That is, the correspondence level between $\mathbf{d}1$ and $\mathbf{d}2$ should be the same as that between $\mathbf{d}2$ and $\mathbf{d}1$.

6) The value allocation of a feature is considered, i.e., the standard deviation of the feature is taken into relation, for its involvement to the similarity between two documents. A feature with a larger spread offers more contribution to the similarity between $\mathbf{d}1$ and $\mathbf{d}2$. For example,

Euclidean does not meet properties 1, 3, 4, and 6, and Cosine, Pairwise-adaptive, Extended Jaccard, Dice, and IT-Sim do not satisfy one or more of properties 3, 4 and 6. Consider three documents $\mathbf{d}1 = < 10, 20 >$ and $\mathbf{d}2 = < 10, 5 >$, and $\mathbf{d}3 = < 10, 0 >$. With Euclidean, the distance between $\mathbf{d}1$ and $\mathbf{d}2$ is 15 which is larger than the distance between $\mathbf{d}2$ and $\mathbf{d}3$, 5. This contradicts properties 1 and 3. With Cosine, the similarity between $\mathbf{d}1$ and $\mathbf{d}2$ is 0.8 which is lower than the similarity between $\mathbf{d}2$ and $\mathbf{d}3$,0.894. This contradicts property 3.

MEASURING THE SIMILARITY BETWEEN THE DOCUMENTS

The relation between the two documents is calculated by using the multi view point technique.The properties 1, 3, 4, and 6, and Cosine, Pairwise-adaptive, Extended Jaccard, Dice, and IT-Sim do not satisfy one or more of properties 3, 4 and 6. Consider three documents $\mathbf{d}1 = < 10, 20 >$ and $\mathbf{d}2 = < 10, 5 >$, and $\mathbf{d}3 = < 10, 0 >$. With Euclidean, the distance between $\mathbf{d}1$ and $\mathbf{d}2$ is 15 which is larger than the distance between $\mathbf{d}2$ and $\mathbf{d}3$, 5. This contradicts properties 1 and 3. With Cosine, the similarity between $\mathbf{d}1$ and $\mathbf{d}2$ is 0.8 which is lower than the similarity between $\mathbf{d}2$ and $\mathbf{d}3$, 0.894. This contradicts property 3. Based on the preferable properties mentioned above, we propose a similarity measure, called SMTP (Similarity Measure for Text Processing), for two documents $\mathbf{d}1 = < d11, d12, \ldots , d1m >$ and $\mathbf{d}2 = < d21, d22, \ldots , d2m >$. Define a function F as follows:

$$F(\mathbf{d}1, \mathbf{d}2) = \frac{\sum_{j=1}^{m} N_*(d_{1j}, d_{2j})}{\sum_{j=1}^{m} N\cup (d_{1j}, d_{2j})} \quad (7)$$

The proposed measure takes into account the following three cases:

a) The similarity value is considered appears in both documents
b) the similarity value is considered appears in only one document
c) the similarity value is considered appears in none of the documents

For the first case, we set a lower bound 0.5 and decrease the similarity as the difference between the feature values of the two documents increases, scaled by a Gaussian function as shown where σj is the standard deviation of all non-zero values for feature wj in the training data set. For the second case, we set a negative constant $-\lambda$ disregarding the magnitude of the non-zero feature value. For the last case, the feature has no contribution to the similarity.

MEASURING THE SIMILARITY VALUES

We extend our method to measure the similarity between two document sets. It is measured by using the TF-IDF measure which is based on the OLP values. The OLP values are the overlapping values which is used to cumulate the similar documents.Refer to Eq.(7), it can be considered as an average score of the features occurring in at least one of the two documents. Based on this perspective, the similarity between two document sets is designed to calculate an average document sets containing q1 and q2 documents, respectively,

i.e., $G1 = \{d^1_1, d^1_2, \ldots, d^1_{q1}\}$ and $G2 = \{d^2_1, d^2_2, \ldots, d^2_{q2}\}$ where $\mathbf{d}^s_j = < d^s_{j1}, d^s_{j2}, \ldots, d^s_{jm} >, s \in \{1,2\}$, and $1 \leq j \leq q1$ or $1 \leq j \leq q2$. The function F between G1 and G2 is defined to be

$$F(G1,G2) = \frac{\sum_{k=1}^{m}\sum_{i=1}^{q1}\sum_{j=1}^{q2} N_*(d^1_{ik}, d^2_{jk})}{\sum_{i=1}^{q1}\sum_{j=1}^{q2}\sum_{k=1}^{q3} N\cup(d^1_{ik}, d^2_{jk})} \quad (8)$$

$$\frac{\sum_{k=1}^{m}\sum_{i=1}^{q1}\sum_{j=1}^{q2} N_*(d^1_{ik}, d^2_{jk})}{\sum_{k=1}^{m}\sum_{i=1}^{q1}\sum_{i=1}^{q2} N\cup(d^1_{ik}, d^2_{jk})} \quad (9)$$

and the similarity measure, S_{SMTP} , for G1 and G2 is

$$S_{SMTP}(G1,G2) = \frac{F(G1,G2)+\lambda}{1+\lambda} \quad (10)$$

The formula Eq.(8) used is essentially an average of similarity values between individual documents. Note that the numerator calculates The similarity between two documents coming from G1and G2, respectively. The numerator is the total sum of these similarity values. The denominator serves the purpose of normalization. With the approximation, a set can be only represented by its center. It is not needed to store the information of the member patterns in a set. Similarity computation with a set is only done with its center instead of with all its member patterns. In general, if the patterns can be nicely clustered, i.e., the patterns of a cluster are close to each other, then the members of the cluster can be nicely represented by the center of the cluster and the quality of the approximation is good.

RESULTS

In this section, we investigate the effectiveness of our proposed similarity measure SMTP using Multi view point technique. By this technique, the related documents forms cluster based on the clustering algorithm. If any additional documents I present, it is either needs to be added in an existing cluster or can form a new cluster based on the hierarchical algorithm. The investigation is done by applying our measure in several text applications, including k-NN based single-label classification (SL-kNN) [18], k-NN based multi-label classification (ML-kNN) [50], k-means clustering (k-means) [9], and hierarchical agglomerative clustering (HAC) [19]. We also compare the performance of SMTP with that of other five measures, Euclidean [45], Cosine [25], Extended Jaccard (EJ) [48],[49], Pairwise-adaptive (Pairwise) [17], and IT-Sim [39], [8]. Note that the percentage of features taken into account for the Pairwise-adaptive measure is set to be 100%. For the Pairwise-adaptive measure, K is determined by the product of the minimum number of non-zero features in the two documents and the percentage of features taken into account. For example, suppose we have two documents **d**1= < 0, 3, 0, 4, 2 > and **d**2 = < 0, 2, 1, 0, 0 >. The minimum number of non-zero features in these two documents is 2. Then we take 2×100% = 2 largest features from **d**1and **d**, respectively. The features from **d**12 are feature 2 and feature 4, while the features from **d**1k are feature 2 and feature 3. The combination of these features contains feature 2, feature 3, and feature 4. In this case, the results obtained by Pairwise-adaptive and Cosine are different. In the following, we use a computer with AMD FX(tm)-4100 Quad-Core Processor 3.6GHz, 8GB of RAM to conduct the experiments. The programming language used is MATLAB7.0.

USES AND APPLICATIONS

A brief description for the four applications is given below.

1) SL-kNN. k-NN [18] is one of the most popular methods for single-label classification in which a document can belong to only one category. It classifies an unseen document by comparing it to its k nearest neighbors in a specified training set. Given a document **d**, let D_k , with corresponding label set L , be a set containing the k most similar documents to **d**.Then **d** is classified to class c which appears most frequently in L_k. A random choice is made when a tie occurs.

2) ML-kNN. ML-kNN [50] is an adaptation of k-NN for multi-label classification in which a document can belong to more than one category. An unseen document is labeled based on its k nearest neighbors using the maximum. For a document **d**, let Dk, with corresponding label set Lk, be a set containing the k most related documents to **d**. If the probability that **d** belongs to class c given Lk is greater than the probability that **d** does not belong to class c given Lk, then **d** is classified to class c.

3) k-means. k-means [9] is one of the most trendy methods which produce a single clustering. It requires the number of clusters, k, to be specified in progress. Initially, k clusters are specified. Then each document in the document set is re-assigned based on the relationship between the documents and the k clusters. Then the
k clusters are reorganized Then all the documents in the document set are re-assigned. This process is continues until the k clusters remains same.

4) HAC. HAC [19] produces a series of clusterings of decreasing number of clusters at each step. The first clustering contains as many clusters as the number of documents in the document set, i.e., each cluster contains one distinctive document. Then the second clustering is produced by merging two most similar clusters into one. This process continues until the final clustering is obtained, which contains only one cluster consisting of all the documents in the document set.

CUMULATIVE DOCUMENTS

In this case, all the relevant documents are collected to form the clusters. The documents are collected based on the similarity values of the documents which is calculated based on the overlapping values. Three data sets, named WebKB [2], Reuters-8 [1], and RCV1 [38], respectively, are used in the experiments presented below. Some important characteristics of the three data sets. Each data set is briefly described below.

1) WebKB. The documents in the WebKB data set are webpages collected by the World Wide Knowledge Base (Web→Kb) project of the CMU

text learning group [13],[43]. The documents were manually classified into several different classes. The data set can be obtained from [2]. The documents of this data set were not predesignated as training or testing patterns. We divide them randomly into training and testing subsets. Among the 4199 documents, 2803 are randomly selected for training and the rest, 1396, are for testing. Table 2 shows the distribution of the documents in each class randomly selected for training and testing, respectively. The number of features involved is 7786.

2) Reuters-8. Reuters-21578 ModeApt`e Split Text Categorization Test Collection [3] contains thousands of documents collected from Reuters newswire in 1987. The most widely used version is Reuters-21578 ModeApt`e, which contains 90 categories and 12902 documents. We use the 8 most frequent ones of the 90 categories and all the documents with less than or more than one topic are removed. The resulting data set is named Reuters-8 in which about 71% (5485/7674) of the documents were predesignated for training and the other documents, about 29% (2189/7674), were predesignated for testing. The distribution of the documents in each class for training and testing. The data set can be obtained from [1]. The number of features involved is 17745.

3) The RCV1 data set consists of 804414 news stories produced by Reuters from 20 Aug 1996 to 19 Aug 1997. There are 47236 features and 101 categories involved in this data set. We use 5 subsets of topics in LYRL2004 split defined in Lewis et al. [38]. The data set we use contains 30000 documents, of which 15000 were predesignated for training and the rest were predesignated for testing. The 5 subsets are arbitrarily named Subset1∼Subset5. Each subset has 3000 training patterns and 3000 testing patterns. In PMC denotes the percentage of documents belonging to more than one category and each training document is assigned to 3.176 categories on average.

CLUSTER FORMATION

Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often proximity according to some defined distance measure. Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. The computational task of classifying the data set into k clusters is often referred to as k-clustering. Besides the term data clustering (or just clustering), there are a number of terms with similar meanings, including cluster analysis, automatic classification, numerical taxonomy, botryology and typological analysis. Document clustering aims to group, in an unsupervised way, a given document set into clusters such that documents within each cluster are more similar between each other than those in different clusters. It is an enabling technique for a wide range of information retrieval tasks such as efficient organization, browsing and summarization of large volumes of text documents. Cluster analysis aims to organize a collection of patterns into clusters based on similarity. Clustering has its root in many fields, such as mathematics, computer science, statistics, biology, and economics. In different application domains, a variety of clustering techniques have been developed, depending on the methods used to represent data, the measures of similarity between data objects, and the techniques for grouping data objects into clusters.For a document corpus with p classes and n documents, we remove the class labels. Then we randomly selected one-third of the documents for training/validation and the remaining for testing. Note that the data for training/validation are separate from the data for testing.

CONCLUSION

We have presented a novel similarity measure between two documents. Several desirable properties are embedded in this measure. For example, the similarity measure is symmetric. The presence or absence of a feature is considered more essential than the difference between the values associated with a present feature. The similarity degree increases when the number of presence-absence feature pairs decreases. Two documents are least similar to each other if none of the features have non-zero values in both documents. Besides, it is desirable to consider the value distribution of a feature for its contribution to the similarity between two documents. The proposed scheme has also been extended to measure the similarity between two sets of documents. To improve the efficiency, we have provided an approximation to reduce the complexity involved in the computation. We have investigated the effectiveness of our proposed measure by applying it in k-NN based single-label classification, k-NN based multi-label classification, k-means clustering, and hierarchical agglomerative clustering (HAC) on several real-world data sets. The results have shown that the performance obtained by the proposed measure is better than that achieved by other measure.

REFERENCES

[1] http://web.ist.utl.pt/ acardoso/datasets/.
[2]http://www.cs.technion.ac.il/ ronb/thesis.html.

[3]http://www.daviddlewis.com/resources/test collections/reuters21578/.
[4] http://www.dmoz.org/.
[5] P. K. Agarwal and C. M. Procopiuc. Exact and approximation algorithms for clustering. *Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 658–667, 1998.
[6] D. W. Aha. Lazy learning: Special issue editorial. *Artificial Intelligence Review*, 11(1-5):7–10, 1997.
[7] G. Amati and C. J. V. Rijsbergen. Probabilistic models of information based on measuring the divergence from randomness. *ACM Transactions on Information Systems*, 20(4):357–389, 2002.
[8] J. A. Aslam and M. Frost. An information-theoretic measure for document similarity. *Proceedings of 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 449–450, 2003.
[9] G. H. Ball and D. J. Hall. A clustering technique for summarizing multivariate data. *Behavioral Science*, 12:153–155, 1967.
[10] D. Cai, X. He, and J. Han. Document clustering using locality preserving indexing. *IEEE Transactions on Knowledge and Data Engineering*, 17(12):1624–1637, 2005.
[11] H. Chim and X. Deng. Efficient phrase-based document similarity for clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20(9):1217 – 1229, 2008.
[12] S. Clinchant and E. Gaussier. Information-based models for ad hoc IR. *Proceedings of 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 234–241, 2010.
[13] M. Craven, D. DiPasquo, D. Freitag, A. K. McCallum, T. M. Mitchell, K. Nigam, and S. Slattery. Learning to extract symbolic knowledge form the world wide web. *Proceedings of 15th National Conference on Artificial Intelligence*, 1998.
[14] I. S. Dhillon, J. Kogan, and C. Nicholas. *Feature Selection and Document Clustering.* In Berry MW Ed. A Comprehensive Survey of Text Mining, 2003.
[15] I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1):143–175, 2001.
[16] I. S. Dhillon, S. Mallela, and R. Kumar. A Divisive infromation-theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Research*, 3:1265–1287, 2003.
[17] J. D'hondt, J. Vertommen, P.-A. Verhaegen, D. Cattrysse, and J. R. Duflou. Pairwise-adaptive dissimilarity measure for document clustering. *Information Sciences*, 180:2341–2358, 2010.
[18] R. O. Duda, P. E. Hart, and D. J. Stork. *Pattern Recognition.* Wiley, 2001.
[19] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Science*, 95(25):14863– 14868, 1998.
[20] H.'Fang, T. Tao, C. Zhai. A formal study of heuristic retrieval constraints. *Proceedings of 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–56, 2004.

