Semantic based Sentence Ordering with Ontology-Mining in Heterogeneous Data using Explicit Semantic Analysis

Kiruthika. P
M.E CSE-PG Student
Department of CSE, Builders Engineering College,
Nathakadaiyur, Kangayam, Tiruppur, Tamilnadu

Satheesh Kumar. A
MCA., M.E, Assistant Professor,
Department of CSE, Builders Engineering College,
Nathakadaiyur, Kangayam, Tiruppur, Tamilnadu

Abstract: An ontology-oriented architecture where core ontology has been used as knowledge base (KB) and allows data integration of different heterogeneous sources. In existing model used to Natural Language Processing and Artificial Intelligence methods to process and mine data in the health sector to uncover knowledge hidden in diverse data sources. The approach has been applied in the field of personalized medicine (study, diagnosis, and treatment of diseases customized for each patient). AI methods have been used with the objective to mine data in the healthcare sector to uncover knowledge hidden in heterogeneous data sources. A set of learned rules (using Data mining techniques on structured data, DM rules) and their improvements (applying NLP techniques on data from the Web) are obtained. In additionally proposed system, to apply three phase Ontology, first stop word removal, stemming and semantic (Synonym word) replacement is used for preprocessing. Next phase Naïve Bayes classification is used. Next phase Rules Extraction is processed and final phase Explicit Semantic analysis is made. In this method automatically construct and incorporate document and word constraints to support unsupervised constrained clustering. The result of the evaluation demonstrates the superiority of our approaches against a number of existing approaches.

Keywords: ***Data Mining, Ontology Mining, Classification Model, Clustering, Automatic Analysis***.

I. INTRODUCTION

Data mining is the process of extracting keywords from data. It is seen as an increasingly important tool by modern business to transform data into an informational advantage. It ranges of profiling practices, such as marketing, surveillance, fraud detection, and scientific discovery.

It is the process of applying these methods to data with the intention of uncovering hidden patterns. It has been used by businesses, scientists and governments to sift through volumes of data such as airline passenger trip records, census data and supermarket scanner data to produce market research reports. (Note, however, that news isn't forever thought-about to be data processing.)

A. Data Mining Tasks

Data mining commonly involves four classes of tasks:

- Clustering –the task of discovering groups and structures in the data that are in some way or another "similar" ways, without using the known structures in the data.
- Classification -the task of generalizing the known structure to apply to new data. For example, an email program might attempt to classify an email as a legitimate or spam. An algorithm includes decision tree learning, nearest neighbor, naive Bayesian classification, neural networks and support vector machines.
- Regression – It attempts to find a function which models the datas with the least error.
- Association rule learning – It searches for relationships between variables. For example a grocery may gather knowledge on the client buying habits.
- Using association rule learning, the supermarkets can determine which products are frequently bought together and use this information for marketing purposes. This is typically cited as market basket analysis.

B. Objective

Document bunch is being studied from several decades however still it's aloof from a trivial and solved drawback. The challenges are:

- A systematic approach called multi-document summarization is required to generate a summary about particular topic.
- It should address the semantic relationship between the sentences in the summary.
- Since data mining based on health records and diseases has become a hot research topic, Explicit Semantic Analysis should be carried out.
- Rules should be extraction such that Co-occurrence of symptom/medicine/treatment details.
- It should improve the health industry in diagnosis thereby.

II.RELATED WORKS

David Gil and Antonio Fernandez [1]: It describe the Internet of Things (IoT) has made it possible for devices around the world to acquire information and store it, in order to be ready to use it at a later stage. However, this potential opportunity is often not exploited because of the excessively big interval between the data collection and the capability to process and analyze it.We review the current IoT technologies, approaches and models in order to discover what challenges need to be met to make more senses of data. The main goal is to review the surveys related to IoT in order to provide well integrated and context aware intelligent services for IoT. Moreover, we present a state-of-the-art of IoT from the context aware perspective that allows the integration of IoT and social networks in the emerging Social Internet of Things (SIoT) term.

M. Fazio and a. Celesti [2]: It describes monitoring activities detect changes in the environment and can be used for several purposes. To develop new advanced services for sensible environments, information gathered throughout the watching got to be keep, processed and correlated to different pieces of information that characterize or influence the environment itself. We propose a Cloud storage answer able to store large quantity of heterogeneous information, and provide them in a uniform way. To this aim, we adopt hybrid architecture that couple Document and Object oriented strategies, in order to optimize data storage, querying and retrieval. We gift the design style and discuss some implementation details within the development of the design among a selected use case.

Javier Medina And Macarena Espinilla [3]: It describe a data streams generated from devices collecting data from patients, which are monitored within both clinical and home environments, provide useful information for decision making processes. Nevertheless, medical personnel are still required to review and process the data and therefore spend a lot of time and effort to detect situations of concern such as exacerbations with conditions or the occurrence of anomalies in the measurements. We propose a technique for the period linguistic analysis of knowledge streams generated from medical observation devices supported a rule-based reasoning engine exploiting a fuzzy linguistic approach. A case study supported health knowledge provided by the Physiological knowledge Modeling Contest is employed let's say the projected methodology and to demonstrate the flexibleness to interpret, in a linguistic manner, knowledge streams and also the detection of risk things of interest supported linguistic expression.

Rimma pivovarov and Adler j. Perotte [4]: It presents the unsupervised Phenome Model (UPhenome), a probabilistic graphical model for large-scale discovery of machine models of sickness, or phenotypes. We tackle this challenge through the joint modeling of a large set of diseases and a large set of clinical observations. The observations square measure drawn directly from heterogeneous patient record knowledge (notes, laboratory tests, medications, and diagnosis codes), and the diseases are modeled in an unsupervised fashion. We apply UPhenome to 2 qualitatively completely different mixtures of patients and diseases: records of very sick patients within the medical care unit with constant observation, and records of outpatients regularly followed by care providers over multiple years. We demonstrate that the UPhenome model can learn from these different care settings, without any additional adaptation. The results shows that (i) the learned phenotypes combine the heterogeneous data types more coherently than baseline LDA-based phenotypes; (ii) they each represent single diseases rather than a mix of diseases additional typically than the baseline ones; and (iii) once applied to unseen patient records, they are correlated with the patients' ground-truth disorders.Code for coaching, inference, and quantitative analysis is formed accessible to the analysis community.

Cristina Soguero-Ruiz and Kristian Hindberg [5]: It developed a learning system capable of exploiting information sent by longitudinal Electronic Health Records (EHRs) for the prediction of a typical operative complication, conjunction leak (AL), in a very
data-driven method and by fusing temporal population data from altogether completely different and heterogeneous sources among the EHRs.We used linear and non-linear kernel ways separately for every information supply, and investing the powerful multiple kernels for their effective combination. To validate the system, we used data from the EHR of the gastrointestinal department at a university hospital. We 1st investigated the first prediction performance from every knowledge supply severally, by computing Area under the Curve values for processed free text (0.83), blood tests (0.74), and vital Signs (0.65), respectively.

When exploiting the heterogeneous data sources combined exploitation the composite kernel framework, the prediction capabilities increased significantly (0.92).Finally, posterior probabilities were evaluated for risk assessment of patients as an aid for clinicians to boost alertness at associate degree early stage, so as to act promptly for avoiding AL complications.Machine-learning applied mathematics model from EHR knowledge will be helpful to predict surgical complications.The combination of EHR extracted free text, blood samples values, and patient vital signs, improves the model performance. These results will be used as a framework for surgical clinical call support.

Didier Dubois And Weiru Liu [6]: It propose and advocate basic principles for the fusion of incomplete or uncertain information items that should apply regardless of the formalism adopted for representing pieces of information coming from several sources. This formalism are often supported sets, logic, partial orders, risk theory, belief functions or inexact possibilities.We propose a general notion info|of data of knowledge} item representing incomplete or unsure information concerning the values of associate entity of interest. It is speculated to rank such values in terms of relative plausibleness, and expressly signifies not possible values. Basic problems touching the results of the fusion method, like relative info content and consistency of data things, likewise as their mutual consistency, square measure mentioned.

Diogo Machado and Tiago Paiva [7]: It describes Diabetes management is a complex and a sensible problem as each diabetic is a unique case with particular needs. The optimal solution would be a constant monitoring of the diabetic's values and automatically acting accordingly. We propose an approach that guides the user and analyses the data gathered to give individual advice. By using data mining algorithms and methods, we uncover hidden behavior patterns that may lead to crisis situations. These patterns can then be transformed into logical rules, able to trigger in a particular context, and advise the user. We believe that this solution is not only beneficial for the diabetic, but also for the doctor accompanying the situation. The advice and rules are useful input that the medical expert can use while prescribing a particular treatment. During the data gathering phase, when the number of records is not enough to attain useful conclusions, a base set of logical rules, defined from medical protocols, directives and/or advice, is responsible for advise and guiding the user. The proposed system will accompany the user at start with generic advice, and with constant learning, advise the user more specifically. We discuss this approach describing the architecture of the system, its base rules and data mining component. The system is to be incorporated in a currently developed diabetes management application for Android.
In this work two different methods were used: association rules and Bayesian networks.

- Association rules: reveal links, and the weight of these links, between variables. By applying this algorithm to our users we were able to conclude rules such as e.g. "At Wednesday in the afternoon your usual meal results in high glycaemic values.". This connection of days and times of the week with crisis occurrences is fundamental to avoid or correct incorrect behaviors.
- Bayesian networks: show variable probabilistic dependencies. In contrast to the last example, where we found relations between variables, with Bayesian networks it's possible to approach the problem of crisis prediction in a different manner. After creating a network for a user we can now ask e.g. "what is the probability of having hypoglycemia given that today is Thursday."

Behzad Golshan And Alon Halevy [8]: It describes the field of data integration has expanded significantly over the years, from providing a uniform query and update interface to structured databases within an enterprise to the ability to search, exchange, and even update, structured or unstructured data that are within or external to the enterprise. This paper describes the evolution within the landscape of knowledge integration since the work on editing queries victimization views within the mid-1990's.
In addition, we describe two important challenges for the field going forward. The first challenge is to develop smart open- supply tools for various parts of knowledge integration pipelines. The second challenge is to produce practitioners with viable solutions for the long-standing drawback of consistently combining structured and unstructured knowledge.

Maria Ganzha and Marcin Paprzycki [9]: It describes The Internet of Things (IoT) idea, explored across the globe, brings about an important issue: how to achieve inter operability among multiple existing (and constantly created) IoT platforms. In this context, in Gregorian calendar month 2016, the European omission has funded seven projects that are to deal with various aspects of interoperability in the Internet of Things. Among them, the INTER- IoT project is aiming at the design and implementation of, and experimentation with, an open cross-layer framework and associated methodology to provide voluntary interoperability among heterogeneous IoT platforms. While the project considers ability across all layers of the software system stack, we are particularly interested in answering the question: how ontologies and semantic data processing can be harnessed to facilitate interoperability across the IoT landscape. Henceforth, we've engaged in a very "fact finding mission" to determine what's presently at our disposal once linguistics ability thinks about. Since the INTER-IoT project is initially driven by two use cases originating from (i) (e/m) Health and (ii) transportation and logistics, these two application domains were used to provide context for our search. The paper summarizes our findings and provides foundation for developing ways and tools for supporting linguistics ability within the INTER-IoT project (and beyond)

Antoni Olivé [10]: It describes a vision of a universal ontology (UO) aiming at solving, or at least greatly alleviating, the semantic integration problem in the field of conceptual modeling and the understandability problem in the field of the semantic web. So far it's been assumed that the UO isn't possible in follow, but we think that it is time to revisit that assumption in the light of the current state-of-the-art. This paper aims to be a step during this direction.
We create an initial proposal of a possible UO.We gift the scope of the UO, the kinds of its concepts, and the elements that could comprise the specification of each concept. We propose a standard structure for the UO consisting of 4 levels. We argue that the UO wants an entire set of construct composition operators, and that we sketch 3 of them. We also tackle a few issues related to the feasibility of the UO, which we think that they could be surmountable.
Finally, we have a tendency to discuss the desirability of the UO, and we explain why we conjecture that there are already organizations that have the knowledge and resources needed to develop it, and that might need AN interest in its development within the close to future.

III. METHODOLOGY

A.EXISTING WORKS

Document clustering has been investigated for use in a number of different diseases of text mining and information retrieval. Initially, Text bunch was investigated for up the exactness or recall in info retrieval systems associate degreed as an economical manner of finding the closest neighbors of a Text. Text clustering has also been used to automatically generate hierarchical clusters of Texts and then uses these clusters to produce an effective Text

classifier for new Texts. Ontology textual data, one of the most important distance measures is Text similarity. Since Text similarity is often determined by word similarity, the semantic relationships between words may affect Text ontology results.

The sharing common named entities (NE) among Texts can be a cue for ontology these Texts together. Moreover, the relationships among vocabularies such as synonyms, antonyms, heteronyms, and hyponyms, may also affect the computation of Text similarity. Text ontology has been number of analysis for different Text database model such as HTML document, XML document and SGML document. The existing system only investigated for use in a number of different diseases of text mining but the number of different kinds of Text ontology need information retrieval. So improve the Text ontology method for web mining techniques in this thesis work. The proposed system is developed an application for recommendations of news diseases to the readers of a news portal. The following challenges gave us the motivation to use ontology of the news diseases:

- The number of available diseases was large.
- A large number of diseases were added each day.
- Diseases corresponding to same news were added from different sources.
- The recommendations had to be generated and updated in real time.

The ontology algorithm is reducing and search Texts for recommendations in users have been interest to a few numbers of clusters of Texts. This improved our time efficiency to a great extent and different from sources Texts. The main motivation of this work has been to investigate possibilities for the improvement of the effectiveness of Text ontology by finding out the main reasons of ineffectiveness of the already built algorithms and get their solutions.

Initially applied the K-Means and Agglomerative Hierarchical ontology methods on the data and found that the results were not very satisfactory and the main reason for this was the noise in the graph, created for the data. The tried for pre-processing of the graph to remove the extra edges.

Heuristic applied for removing the inter cluster edges and then applied the standard graph ontology methods to get much better results. The information tried a completely different approach by first ontology the words of the Texts by using a standard ontology approach and thus reducing the noise and then using this word cluster to cluster the Texts. Their results are found that this also gave better results than the classical K-Means and ontology algorithm methods**.**

1. Cosine Similarity

In this module, two documents are selected. Then the vector values for two documents are found out. The cosine similarity measure is applied. Then the correlation between two documents is found out using the following formula.

$$Corr(u,v) = \frac{u^T v}{\sqrt{u^T u}\sqrt{v^T v}} = \left\langle \frac{u}{\|u\|}, \frac{v}{\|v\|} \right\rangle$$

Correlation Formula

For example, the string “I have to go to school” is present in one document. The string “I have to go to temple” is present in other document. Then the data is prepared such that

[i , have , to , go, school , temple] = [1,1,2,1,1,0]
[i , have , to , go , school , temple] = [1,1,2,1,0,1]
[i , have , to , go , school , temple] = [1,1,2,1,0,1]

Formula: cos = 1*1 + 1*1 + 2*2 + 1*1 + 1*0 + 0*1 / sqrt((1^2 + 1^2 + 2^2 + 1^2 + 1^2 + 0^2) * 1^2 + 1^2 + 2^2 + 1^2 + 1^2 + 0^2)

2. Lexical Unigram, Bigram, Skip-Gram Match

- Two sentences (from text files) are taken. Contents are fetched into strings. Sentences are split into Words.
- For Unigram, the presence of various unigrams in the sentence i is checked against the sentence j.
- For Bigram, Each bigram in the sentence j is searched for their presence in the corresponding sentence i part.
- A skip gram is any combination of words in the order as they appear in a sentence but allowing gap between word occurrences.
- Here, 1-skip bigram is considered where 1-skip bigram allowing one word gap between words in a sentence as they appear.

3. Lexical Longest Common Subsequence

- The longest common subsequence of sentence i – sentence j pair is the longest sequence of words that is common to both sentences.
- Lexical_LCS_match= LCS (sentence i, sentence j)/length of unigrams in the sentence j.
- If the value of Lexical_LCS_match is 0.80 or more, i.e., the length of the common words in pair of sentence is greater than the length of the sentence j, then the sentence combine is taken into account as AN illation combine.

B.PROPOSED WORKS

The proposed system developing an application for recommendations of news diseases to the readers of a news portal.

1. Preprocessing Document

Add Stem Word Document

It enters the given word and stem word using text box control and click save button stem word saved into the table. The details are saved in ‘Stemword’ table. The stem word details view on grid view controls.

Add Stop Word Document

It enters the stop word using text box control and click save button stop word saved into the table. The details are saved in ‘Stopword’ table. The stop word details view on grid view controls.

Add Synonym Word Document

In this module, enter the given word and synonym word using text box control and click save button synonym word saved into the table. The details are saved in 'synonym word' table. The synonym word details view on grid view controls.

2. Document Selection

We can select any document files like Text Files, HTML files and XML files.

3. Process

A. Binary Representation of features (Keywords)

It is the task of classifying the weather of a given set into 2 teams (predicting that cluster every one belongs to) on the premise of a classification rule. Contexts requiring a decision as to whether or not an item has some qualitative property, some specified characteristic, or some typical binary classification include.

Medical testing is to determine if a patient has certain disease or not the classification property is the presence of the disease.

A "pass or fail" check methodology or internal control, i.e deciding if a specification has or has not been met a Go/no go classification.

B.TF-IDF (Term Frequency-Inverse Document Frequency)

It is a numerical statistic that is intended to reflect how important a word is to document in a collection or corpus.It is often used as a weighting factor in searches of information retrieval.

The tf-idf price will increase proportionately to the quantity of times a word seems within the document and is offset by the quantity of documents within the corpus that contain the word.

$$t_f(t,d) = 0.5 + 0.5.\frac{f_{t,d}}{Max\{f_{t',d} : t' \in a\}}$$

The inverse document frequency may be a live of what proportion data the word provides i.e.if it's common or rare across all documents.

$$idf(t,d) = log\frac{N}{|\{d \in D : t \in d\}|}$$

With N: total variety of documents within the corpus N=|D|

$|\{d \in D : t \in d\}|$: Number of documents where the term t appears (i.e., tf (t, d) ≠ 0). If the term isn't within the corpus, this may cause a division –by-zero.

Then tf-idf is calculated as

$$tfidf(t,d,D) = tf(t,d).id$$

C.Naive Bayes Classification

It is a classification technique based on Bayes Theorem with an assumption of Independence among predicatiors.It represents a supervised learning method as well as a statistical method for classification.

$$p(c|x) = \frac{p(x|c)p(c)}{p(x)}$$

- P (c|x) is that the posterior chance of sophistication (c, target) given predictor (x, attributes).
- P(c) is the prior probability of class.
- P (x|c) is that the probability that is that the chance of predictor given category.
- P(x) is the prior probability of predictor

D.Co-Training

It is a machine Learning algorithmic rule used once there area unit solely little amounts of labeled information and enormous amounts of unlabeled information one among its uses is in text mining for search engines.

It is associate semi-supervised learning techniques that needs 2 views of the info.Each example is described using two different feature sets that provides different complementary information about the instance.

E.Clustering [k-Means]

Given a collection of observations ($x_1,x_2,\ldots x_n$),where each observation is a d-dimensional real vector,k-means clustering aims to partition the n observations into k(≤n) sets S={$S_1,S_2,\ldots S_k$} so as to minimize the within cluster sum of squares(WCSS).Formally,the objective is to find:

$$args_s min \sum_{i=1}^{k}\sum_{x\in s_i}\left|\left|X-\mu_i\right|\right|^2 = arg_s min \sum_{i=1}^{k}|s_i|VarS_i$$

Where µi is that the mean of points in Si.This is adored minimizing the pairwise square deviations of points within the same cluster:

$$\text{args}_s\text{min}\sum_{i=1}^{k}\frac{1}{2|s_i|}\sum_{X,Y\in s_i}||X-Y||^2$$

The equivalence can be deduced from identity

$$\sum_{X\in S_i}||X-Y||^2 = \sum_{X\neq Y\in s_i}(X-\mu_i)(\mu_i-Y)$$

4. Rule Extraction Ontology

- Co-occurrence of medicine names with disease in single paragraph is retrieved during rules extraction.
- Co-occurrence of medicine names, disease and treatment in single paragraph is retrieved during rules extraction.

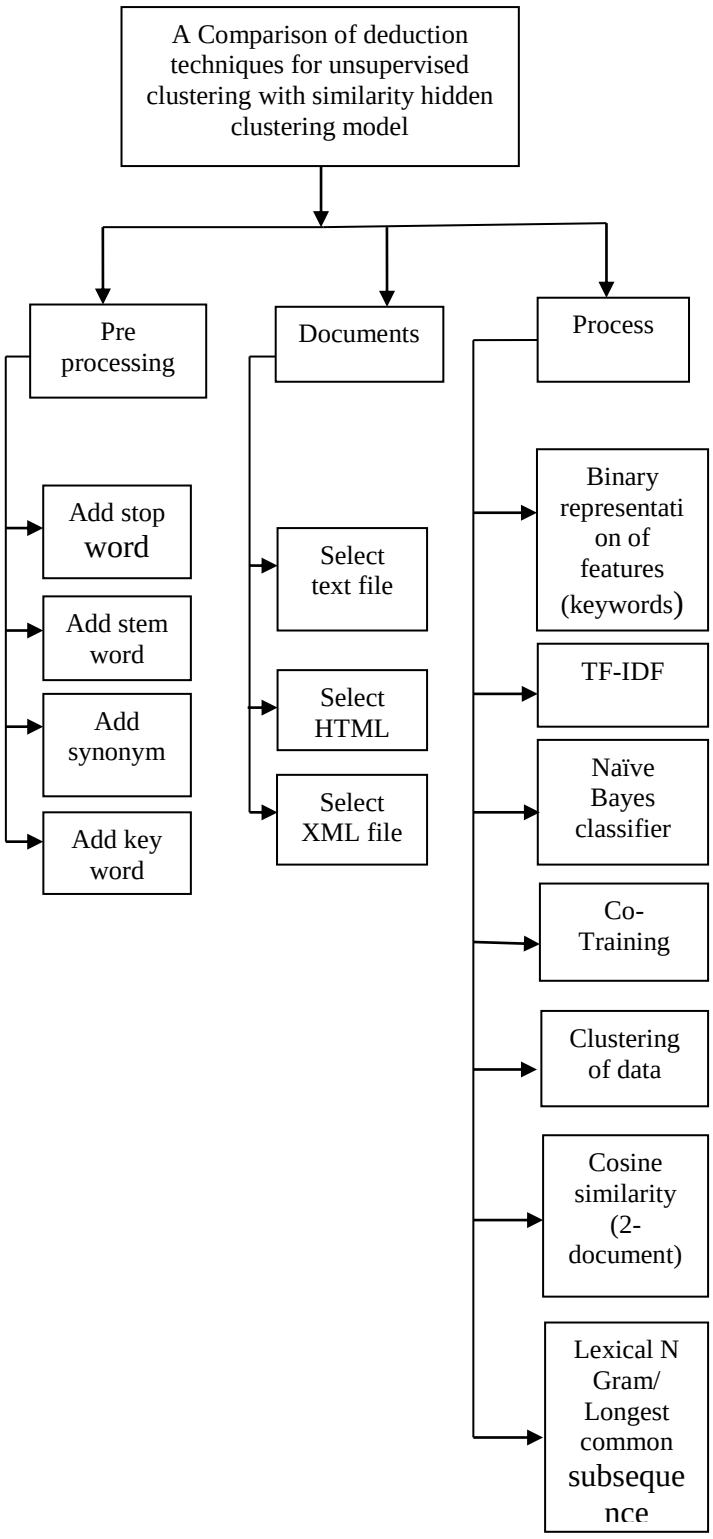

Fig.1 The Block Diagram for Proposed Work

IV CONCLUSION

This survey demonstrated how to construct various text and word constraints and apply them to the constrained co-ontology process. A novel constrained co ontology approach is proposed that automatically incorporates various word and document constraints into information-theoretic co- ontology. It demonstrates the effectiveness of the proposed method for clustering textual documents. There are several directions for future research. The current investigation of unsupervised constraints remains. Furthermore, the algorithm consistently outperformed all the tested constrained clustering and co-clustering methods under different conditions. The enhanced cosine similarity approach results in better ontology process. The future enhancements may be created for documents of various languages. Investigation for higher text options which will be mechanically derived by exploitation linguistic communication process or data extraction tools may be created.

REFERENCES

[1] A. Jain, M. Murty, and P. Flynn, "Data Clustering: A Review," ACM Computing Surveys, vol. 31, no. 3, pp. 264-323, 1999.
[2] I.S. Dhillon, "Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning," Proc. Seventh ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 269-274, 2001.
[3] I.S. Dhillon, S. Mallela, and D.S. Modha, "Information-Theoretic Co-Clustering," Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 89-98, 2003.
[4] H. Cho, I.S. Dhillon, Y. Guan, and S. Sra, "Minimum Sum-Squared Residue Co-Clustering of Gene Expression Data," Proc. Fourth SIAM Int'l Conf. Data. Mining (SDM), 2004.
[5] F. Wang, T. Li, and C. Zhang, "Semi-Supervised Clustering via Matrix Factorization," Proc. SIAM Int'l Conf. Data. Mining (SDM), pp. 1-12, 2008.
[6] Y. Chen, L. Wang, and M. Dong, "Non-Negative Matrix Factorization for Semi-Supervised Heterogeneous Data Co- Clustering," IEEE Trans. Knowledge and Data Eng., vol. 22, no. 10, pp. 1459-1474, Oct. 2010.
[7] R.G. Pensa and J.-F. Boulicaut, "Constrained Co-Clustering of Gene Expression Data," Proc. SIAM Int'l Conf. Data Mining (SDM), pp. 25-36, 2008.
[8] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D.S. Modha, "A Generalized Maximum Entropy Approach to Bregman Co- Clustering and Matrix Approximation," J. Machine Learning Research, vol. 8, pp. 1919-1986, 2007.
[9] K. Nigam, A.K. McCallum, S. Thrun, and T.M. Mitchell, "Text Classification from Labeled and Unlabeled Documents using EM," Machine Learning, vol. 39, no. 2/3, pp. 103-134, 2000.

[10] K. Wagstaff, C. Cardie, S. Rogers, and S. Schro ¨dl, "Constrained K- Means Clustering with Background Knowledge," Proc. 18th Int'l Conf. Machine Learning (ICML), pp. 577-584, 2001.
[11] G. Salton and M. J. McGill. Introduction to Modern Retrieval. McGraw-Hill Book Company, 1983.
[12] E. M. Voorhees. The effectiveness and efficiency of agglomerative hierarchic clustering in document retrieval. PhD thesis, Cornell University, 1986.
[13] I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. Machine Learning, 42(1):143–175, January 2001. Also appears as IBM Research Report RJ 10147, July 1999.
[14] R. V. Katter. Study of document representations: Multidimensional scaling of indexing terms. System Development Corporation, Santa Monica, CA, 1967.
[15] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In ACM SIGIR, 1992.
[16] C. J. Crouch. A cluster-based approach to thesaurus construction. In ACM SIGIR, pages 309–320, 1988.
[17] A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs, NJ, 1988.
[18] J. A. Hartigan. Direct clustering of a data matrix. Journal of the American Statistical Association, 67(337):123–129, March 1972.
[19] I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. Machine Learning, 42(1):143–175, January 2001.
[20] Y. Cheng and G. Church. Biclustering of expression data. In Proceedings ISMB, pages 93–103. AAAI Press, 2000.