Incremental Mining from News Streams

Seokkyung Chung*
Department of Computer Science and Integrated Media System Center
University of Southern California
Los Angeles, California 90089-0781
USA
voice: +1 213-740-4521
fax: +1 213-740-5807
email: seokkyuc@usc.edu

Jongeun Jun
Department of Computer Science and Integrated Media System Center
University of Southern California
Los Angeles, California 90089-0781
USA
voice: +1 213-740-6502
email: jongeunj@usc.edu

Dennis McLeod
Department of Computer Science and Integrated Media System Center
University of Southern California
Los Angeles, California 90089-0781
USA
voice: +1 213-740-4504
email: mcleod@usc.edu

(* Corresponding author)

Incremental Mining from News Streams

Seokkyung Chung, Jongeun Jun, and Dennis McLeod

University of Southern California, USA

INTRODUCTION

With the rapid growth of the World Wide Web, Internet users are now experiencing overwhelming quantities of online information. Since manually analyzing the data becomes nearly impossible, the analysis would be performed by automatic data mining techniques to fulfill users' information needs quickly.

On most Web pages, vast amounts of useful knowledge are embedded into text. Given such large sizes of text collection, mining tools, which organize the text datasets into structured knowledge, would enhance efficient document access. This facilitates information search and at the same time, provides an efficient framework for document repository management as the number of documents becomes extremely huge.

Given that the Web has become a vehicle for the distribution of information, many news organizations are providing newswire services through the Internet. Given this popularity of the Web news services, text mining on news datasets has received significant attentions during the past few years. In particular, as several hundred news stories are published everyday at a single Web news site, triggering the whole mining process whenever a document is added to the database is computationally impractical. Therefore, efficient incremental text mining tools need to be developed.

BACKGROUND

The simplest document access method within Web news services is keyword-based retrieval. Although this method seems effective, there exist at least three serious drawbacks. First, if a user chooses irrelevant keywords, then retrieval accuracy will be degraded. Second, since keyword-based retrieval relies on the syntactic properties of information (e.g., keyword counting), *semantic gap* cannot be overcome (Grosky, Sreenath, and Fotouhi, 2002). Third, only expected information can be retrieved since the specified keywords are generated from users' knowledge space. Thus, if users are unaware of the airplane crash that occurred yesterday, then they cannot issue a query about that accident even though they might be interested.

The first two drawbacks stated above have been addressed by query expansion based on domain-independent ontologies. However, it is well known that this approach leads to a degradation of precision. That is, given that the words introduced by term expansion may have more than one meaning, using additional terms can improve recall, but decrease precision. Exploiting a manually developed ontology with a controlled vocabulary would be helpful in this situation (Khan, McLeod, and Hovy, 2004). However, although ontology-authoring tools have been developed in the past decades, manually constructing ontologies whenever new domains are encountered is an error-prone and time-consuming process. Therefore, integration of knowledge acquisition with data mining, which is referred to as *ontology learning*, becomes a must (Maedche, and Staab, 2001).

To facilitate information navigation and search on a news database, clustering can be utilized. Since a collection of documents is easy to skim if similar articles are grouped together, if the news articles are hierarchically classified according to their topics, then a query can be formulated while a user navigates a cluster hierarchy. Moreover, clustering can be used to identify and deal with near-duplicate articles. That is, when news feeds repeat stories with minor changes from hour to hour, presenting only the most recent articles is probably sufficient. In particular, a sophisticated incremental hierarchical document clustering algorithm can be effectively used to address high rate of document update. Moreover, in order to achieve rich semantic information retrieval, an ontology-based approach would be provided. However, one of the main problems with concept-based ontologies is that topically related concepts and terms are not explicitly linked. That is, there is no relation between *court-attorney*, *kidnap-police*, etc. Thus, concept-based ontologies have a limitation in supporting a topical search. In sum, it is essential to develop incremental text mining methods for intelligent news information presentation.

MAIN THRUST OF THE CHAPTER

In the following, we will explore text mining approaches that are relevant for news streams data.

Requirements of Document Clustering in News Streams

Data we are considering are high-dimensional, large in size, noisy, and a continuous stream of documents. Many previously proposed document clustering algorithms did not perform well on this dataset due to a variety of reasons. In the

following, we define application-dependent (in terms of news streams) constraints that the clustering algorithm must satisfy.

(1) *Ability to determine input parameters*. Many clustering algorithms require a user to provide input parameters (e.g., the number of clusters), which is difficult to be determined in advance, in particular when we are dealing with incremental datasets. Thus, we expect the clustering algorithm not to need such kind of knowledge.

(2) *Scalability with large number of documents*. The number of documents to be processed is extremely large. In general, the problem of clustering n objects into k clusters is NP-hard. Successful clustering algorithms should be scalable with the number of documents.

(3) *Ability to discover clusters with different shapes and sizes*. The shape of document cluster can be of arbitrary shapes, hence we cannot assume the shape of document cluster (e.g., hyper-sphere in k-means). In addition, the sizes of clusters can be of arbitrary numbers, thus clustering algorithms should identify the clusters with wide variance in size.

(4) *Outliers Identification*. In news streams, outliers have a significant importance. For instance, a unique document in a news stream may imply a new technology or event that has not been mentioned in previous articles. Thus, forming a singleton cluster for the outlier is important.

(5) *Efficient incremental clustering*. Given different ordering of a same dataset, many incremental clustering algorithms produce different clusters, which is an unreliable phenomenon. Thus, the incremental clustering should be robust to the input sequence.

Moreover, due to the frequent document insertion into the database, whenever a new document is inserted, it should perform a fast update of the existing cluster structure.

(6) *Meaningful theme of clusters*. We expect each cluster to reflect a meaningful theme. We define "meaningful theme" in terms of precision and recall. That is, if a cluster (*C*) is about "Turkey earthquake", then all documents about "Turkey earthquake" should belong to *C*, and documents that do not talk about "Turkey earthquake" should not belong to *C*.

(7) *Interpretability of resulting clusters*. A clustering structure needs to be tied up with a succinct summary of each cluster. Consequently, clustering results should be easily comprehensible by users.

Previous Document Clustering Approaches

The most widely used document clustering algorithms fall into two categories: partition-based clustering and hierarchical clustering. In the following, we provide a concise overview for each of them, and discuss why these approaches fail to address the requirements discussed above.

Partition-based clustering decomposes a collection of documents, which is optimal with respect to some pre-defined function (Duda, Hart, and Stork, 2001; Liu, Gong, Xu, and Zhu, 2002). Typical methods in this category include center-based clustering, Gaussian Mixture Model, etc. Center-based algorithms identify the clusters by partitioning the entire dataset into a pre-determined number of clusters (e.g., *k*-means clustering). Although the center-based clustering algorithms have been widely used in document clustering, there exist at least five serious drawbacks. First, in many center-based clustering algorithms, the number of clusters needs to be determined beforehand. Second, the algorithm is sensitive to an initial seed selection. Third, it can model only a

spherical (*k*-means) or ellipsoidal (*k*-medoid) shape of clusters. Furthermore, it is sensitive to outliers since a small amount of outliers can substantially influence the mean value. Note that capturing an outlier document and forming a singleton cluster is important. Finally, due to the nature of an iterative scheme in producing clustering results, it is not relevant for incremental datasets.

Hierarchical (agglomerative) clustering (HAC) identifies the clusters by initially assigning each document to its own cluster and then repeatedly merging pairs of clusters until a certain stopping condition is met (Zhao, and Karypis, 2002). Consequently, its result is in the form of a tree, which is referred to as a *dendrogram*. A dendrogram is represented as a tree with numeric levels associated to its branches. The main advantage of HAC lies in its ability to provide a view of data at multiple levels of abstraction. Although HAC can model arbitrary shapes and different sizes of clusters, and can be extended to the robust version (in outlier handling sense), it is not relevant for news streams application due to the following two reasons. First, since HAC builds a dendrogram, a user should determine where to cut the dendrogram to produce actual clusters. This step is usually done by human visual inspection, which is a time-consuming and subjective process. Second, the computational complexity of HAC is expensive since pairwise similarities between clusters need to be computed.

Topic Detection and Tracking

Over the past six years, the information retrieval community has developed a new research area, called TDT (Topic Detection and Tracking) (Makkonen, Ahonen-Myka, and Salmenkivi, 2004; Allan, 2002). The main goal of TDT is to detect the occurrence of

a novel event in a stream of news stories, and to track the known event. In particular, there are three major components in TDT.

(1) *Story segmentation*. It segments a news stream (e.g., including transcribed speech) into topically cohesive stories. Since online Web news (in HTML format) is supplied in segmented form, this task only applies to audio or TV news.

(2) *First Story Detection (FSD)*. It identifies whether a new document belongs to an existing topic or a new topic.

(3) *Topic tracking*. It tracks events of interest based on sample news stories. It associates incoming news stories with the related stories, which were already discussed before. It can be also asked to monitor the news stream for further stories on the same topic.

Event is defined as "some unique thing that happens at some point in time". Hence, an event is different from a topic. For example, "airplane crash" is a topic while "Chinese airplane crash in Korea in April 2002" is an event. Note that it is important to identify events as well as topics. Although a user is not interested in a flood topic, the user may be interested in the news story on the Texas flood if the user's hometown is from Texas. Thus, a news recommendation system must be able to distinguish different events within a same topic.

Single-pass document clustering (Chung, and McLeod, 2003) has been extensively used in TDT research. However, the major drawback of this approach lies in order-sensitive property. Although the order of documents is already fixed since documents are inserted into the database in chronological order, order-sensitive property implies that the resulting cluster is unreliable. Thus, new methodology is required be in terms of incremental news article clustering.

Dynamic Topic Mining

Dynamic topic mining is a framework that supports the identification of meaningful patterns (e.g., events, topics, and topical relations) from news stream data (Chung, and McLeod, 2003). To build a novel paradigm for an intelligent news database management and navigation scheme, it utilizes techniques in information retrieval, data mining, machine learning, and natural language processing.

In dynamic topic mining, a web crawler downloads news articles from a news Web site on a daily basis. Retrieved news articles are processed by diverse information retrieval and data mining tools to produce useful higher-level knowledge, which is stored in a content description database. Instead of interacting with a Web news service directly, by exploiting the knowledge in the database, an information delivery agent can present an answer in response to a user request (in terms of topic detection and tracking, keyword-based retrieval, document cluster visualization, etc). Key contributions of the dynamic topic mining framework are development of a novel hierarchical incremental document clustering algorithm, and a topic ontology learning framework.

Despite the huge body of research efforts on document clustering, previously proposed document clustering algorithms are limited in that it cannot address special requirements in a news environment. That is, an algorithm must address the seven application-dependent constraints discussed before. Toward this end, the dynamic topic mining framework presents a sophisticated incremental hierarchical document clustering algorithm that utilizes a neighborhood search. The algorithm was tested to demonstrate the effectiveness in terms of the seven constraints. The novelty of the algorithm is the

ability to identify meaningful patterns (e.g., news events, and news topics) while reducing the amount of computations by maintaining cluster structure incrementally.

In addition, to overcome the lack of topical relations in conceptual ontologies, a topic ontology learning framework is presented. The proposed topic ontologies provide interpretations of news topics at different levels of abstraction. For example, regarding to a Winona Ryder court trial news topic (*T*), the dynamic topic mining could capture "winona, ryder, actress, shoplift, beverly" as specific terms describing *T* (i.e., the specific concept for *T*) while "attorney, court, defense, evidence, jury, kill, law, legal, murder, prosecutor, testify, trial" as general terms representing *T* (i.e., the general concept for *T*). There exists research work on extracting hierarchical relations between terms from a set of documents (Tseng, 2002). However, the dynamic topic mining framework is unique in that the topical relations are dynamically generated based on incremental hierarchical clustering rather than based on human defined topics such as Yahoo directory.

FUTURE TRENDS

There are many future research opportunities in news streams mining. First, although a document hierarchy can be obtained using unsupervised clustering, as shown in (Aggarwal, Gates, and Yu, 2004), the cluster quality can be enhanced if a pre-existing knowledge base is exploited. That is, based on this priori knowledge, we can have some control while building a document hierarchy.

Second, document representation for clustering can be augmented with phrases by employing different levels of linguistic analysis (Hatzivassiloglou, Gravano, and Maganti, 2000). That is, representation model can be augmented by adding *n*-gram

(Peng, and Schuurmans, 2003), or frequent itemsets using association rule mining (Hand, Mannila, and Smyth, 2001). Investigating how different feature selection algorithms affect on the accuracy of clustering results is an interesting research work.

Third, besides exploiting text data, other information can be utilized since Web news articles are composed of text, hyperlinks, and multimedia data. For example, both terms and hyperlinks (which point to related news articles or Web pages) can be used for feature selection.

Finally, a topic ontology learning framework can be extended to accommodating rich semantic information extraction. For example, topic ontologies can be annotated within Protégé (Noy, Sintek, Decker, Crubezy, Fergerson, and Musen, 2001) WordNet tab. In addition, a query expansion algorithm based on ontologies needs to be developed for intelligent news information presentation.

CONCLUSION

Incremental text mining from news streams is an emerging technology as many news organizations are providing newswire services through the Internet. In order to accommodate dynamically changing topics, efficient incremental document clustering algorithms need to be developed. The algorithms must address the special requirements in news clustering such as high rate of document update, or ability to identify event level clusters as well as topic level clusters.

In order to achieve rich semantic information retrieval within Web news services, an ontology-based approach would be provided. To overcome the problem of concept-based ontologies (i.e., topically related concepts and terms are not explicitly linked), topic

ontologies are presented to characterize news topics at multiple levels of abstraction. In sum, coupling with topic ontologies and concept-based ontologies, supporting a topical search as well as semantic information retrieval can be achieved.

REFERENCES

Aggarwal, C. C., Gates, S. C., & Yu, P. S. (2004). On using partial supervision for text categorization. *IEEE Transactions on Knowledge and Data Engineering, 16 (2),* 245-255.

Allan, J. (2002). Detection as multi-topic tracking. *Information Retrieval, 5 (2-3),* 139-157.

Chung, S., & McLeod, D. (2003, November). Dynamic topic mining from news stream data. *ODBASE'03*. Catania, Sicily, Italy, 653-670.

Duda, R.O., Hart, P.E., & Stork D.G. (2001). *Pattern Classification*. New York: Wiley Interscience.

Grosky, W. I., Sreenath, D. V., & Fotouhi, F. (2002). Emergent semantics and the multimedia semantic web. *SIGMOD Record 31 (4),* 54-58.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining (adaptive computation and machine learning)*. Cambridge, MA: The MIT Press.

Hatzivassiloglou, V., Gravano, L., & Maganti, A. (2000, July). An investigation of linguistic features and clustering algorithms for topical document clustering. *ACM SIGIR Conference on Research and Development in Information Retrieval,* SIGIR'00. Athens, Greece, 224-231.

Khan, L., McLeod, D., & Hovy, E. (2004). Retrieval effectiveness of an ontology-based model for information selection. *The VLDB Journal 13 (1)*, 71-85.

Liu, X., Gong, Y., Xu, W. & Zhu, S. (2002, August). Document clustering with cluster refinement and model selection capabilities. *ACM SIGIR International Conference on Research and Development in Information Retrieval,* SIGIR'02. Tampere, Finland, 91-198.

Maedche, A., & Staab, S. (2001). Ontology learning for the Semantic Web. *IEEE Intelligent Systems, 16 (2)*, 72-79.

Makkonen, J., Ahonen-Myka, H., & Salmenkivi, M. (2004). Simple semantics in topic detection and tracking. *Information Retrieval*, *7 (3-4)*, 347-368.

Noy, N.F., Sintek, M., Decker, S., Crubezy, M., Fergerson, R.W., & Musen M.A. (2001). Creating Semantic Web contents with Protégé -2000. *IEEE Intelligent Systems*, *6 (12)*, 60-71.

Peng, F., & Schuurmans, D. (2003, April). Combining naive Bayes and n-gram language models for text classification. *European Conference on IR Research,* ECIR'03. Pisa, Italy, 335-350.

Tseng, Y. (2002). Automatic thesaurus generation for Chinese documents. *Journal of the American Society for Information Science and Technology, 53 (13)*, 1130-1138.

Zhao, Y., & Karypis, G. (2002, November). Evaluations of hierarchical clustering algorithms for document datasets. *ACM International Conference on Information and Knowledge Management,* CIKM'02. McLean, VA, 515-524.

TERMS AND THEIR DEFINITION

Clustering: An unsupervised process of dividing data into meaningful groups such that each identified cluster can explain the characteristics of underlying data distribution. Examples include characterization of different customer groups based on the customer's purchasing patterns, categorization of documents in the World Wide Web, or grouping of spatial locations of the earth where neighbor points in each region have similar short-term/long-term climate patterns.

Dynamic Topic Mining: A framework that supports the identification of meaningful patterns (e.g., events, topics, and topical relations) from news stream data.

First Story Detection: A TDT component that identifies whether a new document belongs to an existing topic or a new topic.

Ontology: A collection of concepts and inter-relationships.

Text Mining: A process of identifying patterns or trends in natural language text including document clustering, document classification, ontology learning, etc.

Topic Detection and Tracking (TDT): Topic Detection and Tracking (TDT) is a DARPA-sponsored initiative to investigate the state of the art for news understanding systems. Specifically, TDT is composed of the following three major components: (1) segmenting a news stream (e.g., including transcribed speech) into topically cohesive stories; (2) identifying novel stories that are the first to discuss a new event; and (3) tracking known events given sample stories.

Topic Ontology: A collection of terms that characterize a topic at multiple levels of abstraction.

Topic Tracking: A TDT component that tracks events of interest based on sample news stories. It associates incoming news stories with the related stories, which were already discussed before or it monitors the news stream for further stories on the same topic.