Dynamic News Extraction Algorithm

Prerna*
Sanjay Singh**
Smriti Sharma***
Monika Jena****

Abstract

Web has become a great source for news information. Various studies have been devoted to news extraction to find relevant and interesting news article from the large database. Database is also day by day updated with news articles. Earlier greedy approach for extraction of news article following a heuristic approach is used. In this paper we are presenting a novel framework for extracting information found in news articles that are issued in large volumes and which cover similar concepts or issues within a given domain using a dynamic approach. The purpose of dynamic news extraction is to provide an easy and effective accessibility of relevant news articles.

Keywords: Extraction, Algorithm, Mining, Dynamic approach, News extraction

Introduction

News articles contain information rich texts and need the means by which this information can be intelligently searched. The early technique of news extraction is based on keywords and indexing/ matching techniques that may be very sophisticated but the approach suffers from limitations. This paper addresses the particular problem of trying to extract news articles that cannot be extracted through the use of the structure of a news article alone, using a Dynamic News Extraction Algorithm.

Dynamic News Extraction Algorithm includes searching news articles within databases which could either be relational stand-alone databases or hyper textually-networked databases like the World Wide Web. It includes locating from a large news articles collection; those articles that fulfill a specified information need.

In the following sections we will discuss a new technology for news extraction process that is based on dynamic extraction of relevant news articles.

*Research Scholar, B. S. Anangpuria Institute of Technology and Management, Faridabad, INDIA
**Research Scholar, Amity University, Noida, INDIA
***Research Scholar, B. S. Anangpuria Institute of Technology and Management, Faridabad, INDIA
****Assistant Professor, Amity School of Computer Sciences, Noida, INDIA

Authors are grateful to the anonymous referees for comments and suggestions for improving the text and contents of the paper. Authors alone are responsible for any errors/mistakes still remaining in the paper.

The rest of this paper is organized as follows. Section 2 briefly introduces the related approach of news extraction. In section 3, we introduce our novel method of Dynamic News Extraction algorithm. Section 4 summarizes the paper and outlines some interesting directions for future research.

Related Work

Earlier, text mining was used to extract relevant news articles by automatically extracting information from different unstructured news articles stored in the database. The key element behind news extraction is to form new facts or new hypotheses that are explored further by more conventional news extraction techniques.

Current research in the area of text mining tackles the problems of text representation, classification, clustering, and information extraction or the search for and modeling of hidden patterns. In this context the selection of characteristics, domain knowledge and domain-specific procedures plays an important role.

Information retrieval systems and text processing systems have been developed which are quite sophisticated and can retrieve documents by specifying attributes or key words. However, in order to be able to define at least the importance of a word within a given document, usually a vector representation is used, where for each word a numerical "importance" value is stored. The predominant approaches based on this idea are the vector space model, the probabilistic model and the logical model.

Text mining or text data mining, the process of finding useful or interesting patterns, models, directions, trends, or rule from unstructured text, is used to describe the application of data mining techniques to automated discovery of knowledge from text. Text mining has been viewed as a natural extension of data mining, sometimes considered as a task of applying same data mining techniques to specific domain. This reflects the fact that advent of text mining relies on the burgeoning field of data mining to a great degree. Text categorization is a conventional classification problem applied to the textual domain. It solves the problem of assigning text content to predefined categories. In the learning stage, the labeled training data are first pre-processed to remove unwanted details and to "normalize" the data. The keyword extraction from the document is done by identifying and summarizing the contents of the document. The common English words are removed using an "ignore-list" of words during the pre processing stage. And a good heuristic is applied for words that occur frequently in documents of the same class.

While searching a news article, the user looks for something that is already known and has been written by someone else. The problem is to push aside all the news articles that currently is not relevant to user's needs in order to find the relevant information. Text mining technique are used to draw news articles using data mining, machine learning, statistics and computational linguistics.

The problem with Knowledge Discovery from Text (KDT) (Kjetil Nørvag, Randi Øyri, 2005) is to extract explicit and implicit concepts and semantic relations among news articles using Natural Language Processing (NLP) techniques. Its aim is to get insights into large quantities of news articles. KDT, while deeply rooted in NLP, draws on methods from statistics, machine learning, reasoning, information extraction, knowledge management, and others for its news articles discovery process. KDT plays a major role in emerging applications, such as Text Understanding.

Starting with a collection of documents, a text mining tool would retrieve a particular news article and preprocess it by checking format and character sets. Then it would go through an analysis phase, sometimes repeating techniques until desired article is extracted. Technologies (Zhongmin Shi, 2003) have been developed and used in the text mining process for news extraction.

Image Mining

Image mining is the concept used to detect unusual patterns and extract implicit and useful data from images stored in the large data bases and is present in Figure 1. Therefore, we can say that image mining deals with making associations between different images from large image databases. Image mining is used in variety of fields like medical diagnosis, space research, remote sensing, agriculture, industries, and also in handling hyper spectral images. Images include maps, geological structures, and biological structures.

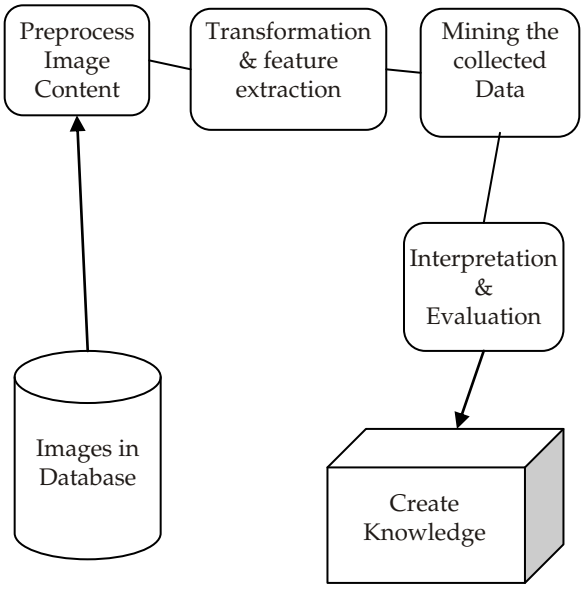

Figure 1: Process of Image Mining

Feature Extraction from Color Images

Image categorization classifies images into semantic databases that are manually re-categorized. In the same semantic databases, images may have large variations with dissimilar visual descriptions (e.g. images of persons, images of industries etc.). In (Raymond Kosala and Hendrik Blockeel, 2000), the authors distinguish three types of feature vectors for image description:

1) Pixel level features,
2) Region level features, and
3) Tile level features.

The Pattern and Knowledge Level is the level that integrates domain related alphanumeric data and the semantic relationships discovered from the image data. Keiji Yanai describes a generic image classification system with an automatic knowledge acquisition mechanism from the Web. Nick Morsillo, Chris Pal, Randal Nelson presented a technique that allows a user to reduce noisy search results and characterize a more precise visual object class. This approach is based on semi-supervised machine learning in a novel probabilistic graphical model made of both generative and discriminative elements.

Framework for mining images by color content is proposed in (Benjamin C. M. Fung, Ke Wang, and Martin Ester, 2003), framework provides the possibility of use distance function for evaluation of similarity among images and 2 type of quantization. The framework focused on color as feature using Color Moment and Block Truncation Coding (BTC) to extract features for image dataset is proposed in (Benjamin C. M. Fung, Ke Wang, and Martin Ester, 2003). Then K-Means clustering algorithm is conducted to group the image dataset into various clusters using Image mining techniques which is based on the Color Histogram, texture of that Image. The query image is taken, then the Color Histogram and Texture is produced and based on this the resultant Image is found. They have investigated histogram based search methods and color texture methods in two different color spaces, RGB and HSV.

Histogram search differentiate an image by its color distribution. It is shown that images retrieved by using the global color histogram may not be semantically associated even though they share similar color distribution in some results. There is need to conduct more research on image mining to see if data mining techniques could be used to classify, cluster, and associate images.

Video Mining

Mining video data is even more complicated than mining image data. One can regard video to be a collection of moving images, much like animation. The important areas include developing query and retrieval techniques for video databases, including video indexing, query languages, and optimization strategies. The strategic view is presented in Figure 2.

In video mining, there are three types of videos:

a) The produced (e.g. movies, news videos, and dramas),
b) The raw (e.g. traffic videos, surveillance videos etc), and
c) The medical video (e.g. ultra sound videos including echocardiogram).

Higher-level information from video includes:

i) Detecting trigger events (e.g. any vehicles entering a particular area, people exiting or entering a particular building),
ii) Determining typical and anomalous patterns of activity, generating person centric or object-centric views of an activity, and
iii) Classifying activities into named categories (e.g. walking, riding a bicycle), clustering and determining interactions between entities.

Shot Detection Methods can be classified into many categories: pixel based, statistics based, transform based, feature based and histogram based, discussed in (Hany Mahgoub, Dietmar Rösner, Nabil Ismail and Fawzy Torkey, 2008). For example, one could examine video clips and find associations between different clips. Other one could find unusual patterns in video clips. Pattern matching in video databases is possible when one has predefined images and then they match these images with the multiple video clips and analyze the video clips. Xie et al have used an unsupervised Hierarchical Hidden Markov Model (HHMM) framework to discover patterns in soccer video. They use low-level features such as motion intensity and dominant color at the lower level of the HHMM and binary labels at the upper level of the HHMM.

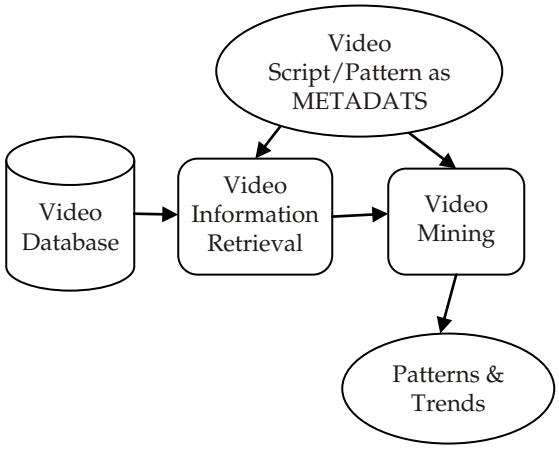

Figure 2: Direct Video Mining

Audio Mining

Since audio is a continuous media type like video, the techniques for audio information processing and mining are similar to video information retrieval and mining. Audio data could be in the form of radio, speech, or spoken language. Even television news has audio data, and in this case audio may have to be integrated with video and possibly text to capture the annotations and captions. To mine audio data, one could convert it into text using speech transcription techniques. Audio data could also be mined directly by using audio information processing techniques and then mining selected audio data. The researchers have used perceptual features such as loudness, brightness, pitch etc for mining the audio.

News Extraction

It includes analysis of unstructured news articles. Key phrases in news articles are identified and relationships within news articles. It is done by looking for predefined sequences in text, using a process called pattern matching.

Summarization

Summarization is immensely helpful for trying to figure out whether or not a lengthy news articles meet the user's needs and is worth reading for further information. With large texts, text summarization software processes and summarizes the document in the time it would take the user to read the first

paragraph. The key to summarization is to reduce the length and detail of a document while retaining its main points and overall meaning.

Categorization

Categorization involves identifying the main themes of a news article by placing the news articles into a pre-defined set of topics. Categorization counts words that appear and, from that count, identifies the main topics that the document covers.

Clustering

Clustering (M. A. Hearst, 2003) is a technique used to group similar news articles. News articles can appear in multiple subtopics, thus ensuring that a useful news article will not be omitted from search results. A basic clustering algorithm creates a vector of topics for each document and measures the weights of how well the news article fits into each cluster.

Proposed Work

Dynamic News Extraction algorithm is based on finding the relevant articles according to the keywords that occur in articles. We find an optimal solution that includes finding the news article following a bottom up approach. It starts with the collection of news article from the database and a set of keywords. Following the user's search keywords are searched in the database of news article.

Dynamc_Best () algorithm is applied that returns a two dimensional array best[][] indicating the number of times keywords are occurring in the article. A key element is considered that provides the minimum condition for keyword to occur in the news article. If the condition for the number of keywords is satisfied in the news article, the frequency of occurrence is set in the array best[][]. The algorithm Dynamic_Best() is given below:

```
Dynamic_Best (Database newsarticle, Keyword key)
1.  {
2.  n= | newsarticle | ;
3.  for key = 0 to key
4.  set best[0,key] = 0;
5.  for i= 1 to n
6.  Set best[i,0]=0;
7.  for i=1 to n
8.  {
9.  for w=1 to key;
10. {
11. If keyi<=key
12. {
13. If ((freqi + best[i-1,key-keyi])>key[i-1,key);
14. set best[i,key]=freqi+best[i-1,key-keyi];
15. Else
16. set best[i,key]= best[i-1,key-keyi];
17. }
18. }
19. Else
20. set best[i,key]= best[i-1,key-keyi];
21. }}
```

Dynamic News Extraction algorithm is applied after obtaining the best[][] array indicating the frequency of keywords in news article. It moves from bottom to top, comparing the frequency of occurrence of keyword in the news article with the frequency of occurrence of another keyword in the same news article. Given below is algorithm Dynamic_News_Extraction ():

```
Dynamic_News_Extraction(Database newsarticle,array best[])
1.  {
2.  n= | newsarticle | ;
3.  set i=n;
4.  set k=key;
5.  while(i>0 and k>0)
6.  {
7.  If best[i,k]!=best[i-1,k]) then
8.  {
9.  Mark the ith newsarticle;
10. set k=k-keyi;
11. set i=i-1;
12. }
13. Else
14. set i=i-1;
15. }
16. Exit
17. }
```

If some news articles give the optimal solution among all news articles in the database, the frequency of all these relevant news article obtained is added to get an optimal solution(in terms of sum of frequency of keywords).

Experiment and Evaluation

User searches for news article, Algorithm Dynamic_Best() returns the frequency of keywords. The final optimal solution set of frequency of keywords in news articles is obtained by applying the Algorithm Dynamic_News_Extraction().

Table 1: News Articles with the frequency of occurrence of keywords in News Article

Keywords / News Articles	Romney	Obama	Nuclear	Baumgartner
US Presidential debate	97	89	2	0
Taliban Attacks	59	0	09	81
Planes bomb rebels in North	81	45	67	31
Supersonic fall test spacesuit	78	78	321	176

If user searches for news on "US Presidential debate", the Dynamic News Extraction algorithm finds the optimal set based on frequency of keywords obtained from Dynamic Best Algorithm as shown:-

Optimal set= {Romney, Obama} = {97+89}
= 186

The overall complexity of Dynamic Extraction algorithm is O(key * n), where, key represents the total number of keywords, and n represents the total number of news articles.

Conclusion

Dynamic News Extraction algorithm is more efficient as compared to other news extraction algorithms. This algorithm follows dynamic approach in bottom up manner to obtain the best and appropriate news article. Further work in this field includes measure of optimality of dynamic extraction algorithm proposed. Future work also includes the implementation of Dynamic News Extraction algorithm in the field of audio, video and image mining.

References

1. *Mabayoje, M. A., Bajeh, A. O. and Olabiyisi, S. O. (2011), " Dynamic Information Extraction (die) Algorithm for Knowledge Accessibility and Reusability in Tertiary Schools", in Proceedings of the International Journal of Computer Information Systems, Vol. 2, No. 4.*
2. *Biswajit Bhowmik (2010), "Dynamic Programming – Its Principles, Applications, Strengths, and Limitations", in Proceedings of the International Journal of Engineering Science and Technology Vol. 2(9), 4822-4826.*
3. *Zhongmin Shi (2003), "Performance Improvement for Frequent Term-based Text Clustering Algorithm", Technique report on Computing Science, Simon Fraser University, April.*
4. *Agrawal, R., Srikant R. (1994), Fast Algorithms for Mining Association Rules in Large Databases, in Proceedings of the VLDB 94, Santiago de Chile, Chile, pp. 487-499.*
5. *Ji-Rui, LI, Kai YANG (2010), "News clustering System Based on Text Mining", in Proceedings of the Advanced Management Science (ICAMS), IEEE International Conference, July.*
6. *Kjetil Nørvag, Randi Øyri (2005), "News Item Extraction for Text Mining in Web Newspapers" in Proceedings of the International Workshop on Challenges in Web Information Retrieval and Integration (WIRI'05).*
7. *Hearst, M. A. (2003), "What is text mining?", http://www.sims.berkeley.edu /~heast/text-mining.html, Oct.*
8. *Karypis, G., Eui-Hong (Sam) Han and Vipin Kumar (1999), "Chameleon: Hierarchical Clustering Algorithm Using Dynamic Modeling", Computer Magazine, August.*
9. *Piatesky-Shapiro G. and Frawley W. (Eds.) (1991), "Knowledge Discovery in Databases", AAAI Press, Menlo Park, CA.*
10. *Hany Mahgoub, Dietmar Rösner, Nabil Ismail and Fawzy Torkey (2008), "A Text Mining Technique Using Association Rules Extraction", in proceedings of the International Journal of Information and Mathematical Sciences, 4:1.*
11. *Benjamin C. M. Fung, Ke Wang, and Martin Esster (2003), "Hierarchical Document Clustering Using Frequent Itemsets", in Proceedings of the 2003 SIAM International Conference on Data Mining (SDM'03), San Francisco, CA, May 1-3, pp. 59-70.*
12. *Florian Beil, Martin Ester, and Xiaowei Xu, "Frequent Term-Based Text Clustering", in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA.*
13. *Raymond Kosala and Hendrik Blockeel (2000), "Web Mining Research: A survey", SIGKDD Exploration, Vol.2 issue 1, July, pp- 1-15.*
14. *Aura Conci. And Everest Mathias M. M. Castro, "Image Mining By Color Content".*
15. *Zhang Ji, Wynne Hsu and Mong Li Lee (2001), "Image Mining: Issues, Frameworks and Techniques", in Proceedings of the 2nd International Workshop on Multimedia Data Mining (MDM/KDD'2001), San Francisco, CA, USA, pp. 13-20.*
16. *Boresczky, J. S. and Rowe, L. A. (1996), "A Comparison of Video Shot Boundary Detection Techniques", Storage & Retrieval for Image and Video Databases IV, Proceedings SPIE 2670, 1996, pp.170-179.*
17. *Yu, H. and Wolf, W. (1997), "A Visual Search System for Video and Image Databases", in Proceedings IEEE International Conference on Multimedia Computing and Systems, Ottawa, Canada, June, pp. 517–524.*
18. *Liu, H. and Motoda, H. (1998), "Feature Extraction, Construction and Selection: A Data Mining Perspective", Kluwer, 59-98. [7] Ji Zhang Wynne Hsu Mong Li Lee "Image Mining: Issues, Frameworks and Techniques".*
19. *Bingbing Ni, Zheng Song, Shuicheng Yan "Web Image Mining towards Universal Age Estimator".*