Integrated Intelligent Research (IIR)

International Journal of Data Mining Techniques and Applications
Volume: 04 Issue: 02 December 2015 Page No.70-74
ISSN: 2278-2419

Text Mining with Automatic Annotation from Unstructured Content

R.Priya[1], R. Padmajavalli[2]

[1]Research Scholar, Research & Development Centre, BharathiarUniversity, Coimbatore ,Assistant Professor in Dept of Computer Science, GSS Jain College for Women, Chennai

[2]Research Supervisor, Research & Development Centre, BharathiarUniversity, Coimbatore, Associate Professor in Dept of Computer Science, Bhaktavatsalam Memorial College for Women,Chennai

Email: rpriyaphd@gmail.com , padmahari2002@yahoo.com

Abstract-Text mining is vast area as compared to information retrieval. Typical text mining tasks include document classification, document clustering, building ontology, sentiment analysis, document summarization, Information extraction etc. Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. Information Extraction is a vital area in Text Mining Techniques, which is an automatic/semi automatic extraction of structured information from unstructured documents. In most of the cases this activity concerns processing human language texts by means of natural language processing (NLP). In this paper, we present themultimedia document processing and automatic annotation out of images/video as information extraction.

keywords—Text Mining,Multimedia Mining,Automatic Image Annotation

I. INTRODUCTION

A multimedia database system stores and manages a large collection of *multimedia data*, such as audio, video, image, graphics, speech, text, document, and hypertext data, which contain text, text markups, and linkages. Multimedia database systems are increasingly common owing to the popular use of audio video equipment, digital cameras, CD-ROMs, and the Internet. Typical multimedia database systems include NASA's EOS (Earth Observation System), various kinds of image and audio-video databases, and Internet databases [1].
Annotatingmultimedia content with semantic informationsuch as scene/segment structures and metadataabout visual/auditory objects is necessary foradvanced multimedia content services [2].

II. SEARCHING MULTIMEDIA DATA

When searching for similarities in multimedia data, we can search on either the data description or the data content. For similarity searching in multimedia data, we consider two main families of multimedia indexing and retrieval systems: (1) Description-based retrieval systems, which build indices and perform object retrieval based on image descriptions, such as keywords, captions, size, and time of creation; and (2) Content-based retrieval systems, which support retrieval based on the image content, such as color histogram, texture, pattern, image topology, and the shape of objects and their layouts and locations within the image [1].

A. Description-based retrieval

Description-based retrieval is labor-intensive it performed through manually or automatic, if performed automatically the results are typically of poor quality. For example, the assignment of keywords to images can be a tricky and arbitrary task. Recent development of Web-based image clustering and classification methods has improved the quality of description-based Web image retrieval, because image surrounded text information as well as Web linkage information can be used to extract proper description and group images describing a similar theme together.

B. Content-based retrieval

Content-based retrieval uses visual features to index images and promotes object retrieval based on feature similarity, which is highly desirable in many applications. In a content-based image retrieval system, there are often two kinds of queries: image sample-based queries and image feature specification queries. Image-sample-based queries find all of the images that are similar to the given image sample. This search compares the feature vector (or signature) extracted from the sample with the feature vectors of images that have already been extracted and indexed in the image database. Based on this comparison, images that are close to the sample image are returned. Image feature specification queries specify or sketch image features like color, texture, or shape. which are translated into a feature vector to be matched with the feature vectors of the images in the database. Content-based retrieval has wide applications, including medical diagnosis, weather prediction, TV production, Web search engines for images, and e-commerce [1].

a) Color histogram–based signature
In this approach, the signature of an image includes color histograms based on the color composition of an image regardless ofits scale or orientation.

b) Multi feature composed signature
In this approach, the signature of an image includes a composition of multiple features: color histogram, shape, image topology, and texture. The extracted image features are stored as metadata, and images are indexed based on such metadata.

c) Wavelet-based signature
This approach uses the dominant wavelet coefficients of an image as its signature. Wavelets capture shape, texture, and

image topology information in a single unified framework. This improves efficiency and reduces the need for providing multiple search primitives [1].

III. AUDIO AND VIDEO DATA MINING

Besides still images, an incommensurable amount of audiovisual information is becoming available in digital form, in digital archives, on the World Wide Web, in broadcast datastreams, and in personal and professional databases. This amount is rapidly growing. There are great demands for effective content-based retrieval and data mining methodsfor audio and video data. Typical examples include searching for and multimedia editing of particular video clips in a TV studio, detecting suspicious persons or scenes in surveillancevideos, searching for particular events in a personal multimedia repository such as My Life Bits, discovering patterns and outliers in weather radar recordings, and finding a particular melody or tune in your MP3 audio album. To facilitate the recording, search, and analysis of audio and video information from multimedia data, industry and standardization committees have made great strides toward developing a set of standards for multimedia information description and compression. For example, MPEG-*k* (developed by MPEGand JPEG are typical video compression schemes[1].

IV. PROCESS OF MULTIMEDIA MINING

The process of applying multimedia mining in order to retrieve different types of data. Data collection is the first and foremost point of a learning system, as the quality of raw data is the factor which determines the overall achievable performance. The main aim of data pre-processing is to discover the important patterns from the raw data, which includes the concepts of data cleaning, normalization, transformation, feature selection etc… Learning can be of straightforward, if informative features can be identified at pre-processing stage. Detailed procedure depends highly on the nature of raw data and problem's domain. The product of data pre-processing is the training set. Given a training set, a learning model has to be chosen to learn from it and make multimedia mining model more iterative. Higher complexity found on compared data mining with multimedia mining: a) the huge volume of data, b) the variability and heterogeneity of the multimedia data (e.g. diversity of sensors, time or conditions of acquisition etc) and c) the multimedia content's meaning is subjective [3].

V. APPLICATIONS OF MULTIMEDIA MINING

The video and audio data mining can be found in the Mining Cinematic Knowledge project [4], which created a movie mining system by examining the suitability of existing concepts in data mining to multimedia.

VI. MULTIMEDIA CONTENT ANNOTATION

Multimedia content annotation is an extension of documentannotation such as GDA (Global DocumentAnnotation). Since naturallanguage text is more tractable and meaningfulthan binary data of visual (image andmoving picture) and auditory (sound and voice)content, we associate text with multimedia contentin several ways. Since most video clipscontain spoken narrations, our system convertsthem into text and integrates them into videoannotation data. The text in the multimediaannotation is linguistically annotated based onGDA [2].

A. Automatic Video Annotation

The linguistic annotation technique has an important role in multimedia annotation. Our video annotation consists of creation of text data related to video content, linguistic annotation of the text data, automatic segmentation of video, semi-automatic linking of video segments with corresponding text data, and interactive naming of people and objects in video scenes. To be more precise, video annotation is performed through the following three steps. First, for each video clip, the annotation system creates the text corresponding to its content. We developed a method for creation of voice transcripts using speech recognition engines.It is called multilingual voice transcription and described later.Second, some video analysis techniques are applied to characterization of visual segments and individual video frames [2].

B. Automatic Image Annotation

Automatic image annotation (also known as automatic image tagging or linguistic indexing) is the process by which a computer system automatically assigns metadata in the form of captioning or keywords to a digital image. This application of computer vision techniques is used in image retrieval systems to organize and locate images of interest from a database [5].This method can be regarded as a type of multi-class image classification with a very large number of classes - as large as the vocabulary size. Typically, image analysis in the form of extracted feature vectors and the training annotation words are used by machine learning techniques to attempt to automatically apply annotations to new images. The first methods learned the correlations between image features and training annotations, then techniques were developed using machine translation to try to translate the textual vocabulary with the 'visual vocabulary', or clustered regions known as *blobs*. Work following these efforts has included classification approaches, relevance models and so on.The advantages of automatic image annotation versus content-based image retrieval (CBIR) are that queries can be more naturally specified by the user [1]. CBIR generally (at present) requires users to search by image concepts such as color and texture, or finding example queries. Certain image features in example images may override the concept that the user is really focusing on. The traditional methods of image retrieval such as those used by libraries have relied on manually annotated images, which is expensive and time-consuming, especially given the large and constantly growing image databases in existence [5].

Given a query word, this model can be used to rank the images using a language modeling approach [10, 11, 12, 15]. While this model is useful for ranked retrieval, it is less useful for people to look at. Fixed length annotations can be generated by using the top N (N = 3, 4 or 5) words to annotate the images. This model is called the fixed annotation-based cross-media relevance model (FAM) [9].

Figure 1: Images “Day” and “Night” automatically annotated

Figure 1 illustrates the power of the relevance model. TheFigure shows two images (from the test set) which were annotated by FAM. Although the two are clearly pictures, the word \Day" was missing from the manual annotations. In these cases, the model allows us to catch errors in manual annotation A natural way to annotate an image I would be to sample n words w1 . . . wn from its relevance model P(·|I). In order to do that, we need to know the probability of observing any given word w when sampling from P(·|I). That is, we need to estimate the probability P(w|I) for every word w in the vocabulary. Given that P(·|I) itself is unknown, the probability of drawing the word w is best approximated by the conditional probability of observing w given that we previously observed b1 . . . bm as a random sample from the same distribution:

$$P(w|I) \approx P(w|b1 \dots bm) \quad (1)$$

We cannot use the prevalent maximum-likelihood estimator for that probability because the image representation b1 . . . bk does not contain any words. However, we can use the training set T of annotated images to estimate the joint probability of observing the word w and the blobs b1 . . . bm in the same image, and then marginalizing the distribution with respect to w. The joint distribution can be computed as the expectation over the images J in the training set:

$$P(w, b1, \dots, bm) = X\ J \in T\ P(J)P(w, b1, \dots, bm|J) \quad (2)$$

We assume that the events of observing w and b1,…. ,bm are mutually independent once we pick the image J, and identically distributed according to the underlying distribution P(·|J). This assumption follows directly from our earlier decision to model each image as an urn containing both words and blobs. Since the events are independent, we can rewrite equation (2) as follows:

$$P(w, b1, \dots, bm) = X\ J \in T\ P(J)P(w|J)\ Ym\ i=1\ P(bi|J) \quad (3)$$

The prior probabilities P(J) can be kept uniform over all images in T . Since the images J in the training set contain both words and blobs, we can use smoothed maximumlikelihood estimates for the probabilities in equation (3). Specifically, the probability of drawing the word w or a blob b from the model of image J is given by:

$$P(w|J) = (1 - \alpha J)\ \#(w, J)\ /|J| + \alpha J\ \#(w, T)\ /|T| \quad (4)$$
$$P(b|J) = (1 - \beta J)\ \#(b, J)\ /|J| + \beta J\ \#(b, T)\ /|T| \quad (5)$$

Here, #(w, J) denotes the actual number of times the word w occurs in the caption of image J (usually 0 or 1, since the same word is rarely used multiple times in a caption). #(w, T) is the total number of times w occurs in all captions in the training set T . Similarly, #(b, J) reflects the actual number of times some region of the image J is labeled with blob b, and #(b, T) is the cumulative number of occurrences of blob b in the training set. |J| stands for the aggregate count of all words and blobs occurring in image J, and |T | denotes the total size of the training set. The smoothing parameters αJ and βJ determine the degree of interpolation between the maximum likelihood estimates and the background probabilities for the words and the blobs respectively. We use different smoothing parameters for words and blobs because they have very different occurrence patterns: words generally follow a Zipfian distribution, whereas blobs are distributed much more uniformly, due in part to the nature of the clustering algorithm that generates them. The values of these parameters are selected by tuning system performance on the held-out portion of the training set. Image Annotation Equations (1) - (5) provide the machinery for approximating the probability distribution P(w|I) underlying some given image I. We can produce automatic annotations for new images by first estimating the distribution P(w|I) and then sampling from it repeatedly, until we produce a caption of desired length. Or we could simply pick a desired number n of words that have the highest probability under P(w|I) and use those words for the annotation.

C. Models of Image Retrieval

The task of image retrieval is similar to the general ad-hoc retrieval problem. We are given a text query Q = w1 . . . wk and a collection C of images. The goal is to retrieve the images that contain objects described by the keywords w1 . . . wk, or more generally rank the images I by the likelihood that they are relevant to the query. We cannot simply use a text retrieval systems because the images I ∈ C are assumed to have no captions. In the remainder of this section we develop two models of image retrieval. The first model makes extensive use of the annotation model developed in the previous section. The second model does not rely on annotations and instead “translates” the query into the language of blobs.

a) Annotation-based Retrieval Model(PAM)

A simple approach to retrieving images is to annotate each image with a small number of keywords. We could then index the annotations and perform text retrieval in the usual manner. This approach is very straightforward, and, is quite effective for single-word queries. However, there are several disadvantages. First, the approach does not allow us to perform ranked retrieval (other than retrieval by coordination-level matching). This is due to the binary nature of word occurrence in automatic annotations: a word either is or is not assigned to the image, it is rarely assigned multiple times. In addition, all annotations are likely to contain the same number of words, so document length normalization will not differentiate between images. As a result, all images containing some fixed number of the query words are likely to receive the same score. The second problem with indexing annotations is that we must a-priori decide what annotation length is appropriate. The number of words in the annotation has a direct influence on the recall and precision of this system. In general, shorter annotations will lead to higher precision and lower recall, since fewer images will be annotated with any given word. Short annotations are more appropriate for a casual user, who is interested in finding a few relevant images without looking at too much junk. On the other hand, a professional user may be interested in higher recall and thus may need longer annotations. Consequently, it would be challenging to field the retrieval system in a way that would suit diverse users. An alternative to fixed-length annotation is to use probabilistic annotation. we developed a technique that assigns a probability P(w|I) to every word w in the vocabulary. Rather than matching the query against the few top words, we could use the entire probability distribution P(·|I) to score images using a language-modeling approach [10, 11,12,15]. In a language modeling approach we score the documents (images) by the probability that a query would be observed during i.i.d. random sampling from a document (image) language model. Given the query Q = w1 . . . wk, and the image I = {b1 . . . bm}, the probability of drawing Q from the model of I is:

P(Q|I) = Yk j=1 P(wj |I)(6)
where P(wj |I) is computed according to equations (1) - (5).

This model of retrieval does not suffer from the drawbacks of fixed-length annotation and allows us to produce ranked lists of images that are more likely to satisfy diverse users.

b) Direct Retrieval Model(DRM)

The annotation-based model outlined in section (a) effect converts the images in C from the blob-language to the language of words. It is equally reasonable to reverse the direction and convert the query into the language of blobs. Then we can directly retrieve images from the collection C by measuring how similar they are to the blob-representation of the query. The approach we describe was originally proposed by [14] for the task of cross-language information retrieval. We start with a text query Q = w1 . . . wk. We assume that there exists an underlying relevance model P(·|Q), such that the query itself is a random sample from that model. We also assume that images relevant to Q are random samples from P(·|Q) (hence the name relevance model). In the remainder of this section we describe: (i) how to estimate the parameters P(b|Q) of this underlying relevance model, and (ii) how we could rank the images with respect to this model. Estimation of the unknown parameters of the query model is performed using the same techniques used in section 4.1. The probability of observing a given blob b from the query model can be expressed in terms of the joint probability of observing b from the same distribution as the query words w1 . . . wk:
P(b|Q) ≈ P(b|w1 . . wk) = P(b, w1 . . wk) P(w1 .. . wk) (7)
The joint probability P(b, w1 . . . wk) can be estimated as an expectation over the annotated images in the training set, by assuming independent sampling from each image J ∈ T :
P(b, w1, . . . , wk) = X J∈T P(J)P(b|J) Yk i=1 P(wi|J) (8) The probabilities P(b|J) and P(wi|J) can be estimated from equation (5).

The prior probabilities P(J) can be kept uniform, or they can be set to reflect query-independent user preferences for a particular type of image, if such information is available. Ranking. Together, equations (7) and (8) allow us to "translate" the query Q into a distribution P(·|Q) over the blob vocabulary. What remains is to specify how this distribution can be used for effective ranking of images I∈ C. One possibility would be to rank the images by the probability that they are a random sample from P(·|Q), as was suggested for the task of ad-hoc retrieval in [13]. In this paper we opt for a specific case [6] of the more general risk minimization framework for retrieval proposed and developed by [15]. In this approach, documents (images) are ranked according to the negative Kullback-Liebler divergence between the query model P(·|Q) and the document (image) model P(·|I): −KL(Q||I) = X b∈B P(b|Q)log P(b|I) P(b|Q) (9) Here P(b|Q) is estimated using equations (7) and (8), while P(b|I) can be computed directly from equation (5), since every image I ∈ C has a blob representation [7].

Query length	1 word	2 words	3 words	4 words
Number of queries	179	386	178	24
Relevant images	1675	1647	542	67
Avg (PAM)	0.1501	0.1419	0.1730	0.2364
Avg (DRM)	0.1697	0.1642	0.2030	0.2765

VII. EXPERIMENTAL RESULTS

In this section we will discuss details of the dataset used and also show experimental results using the different models. Next section compares the results of the fixed length annotation model(FLM) with the Co-occurrence and Translation Models. This is followed by results on the two retrieval models PAM and DRM. Finally, we show some examples to illustrate different aspects of the models. Dataset :Since our focus in this paper is on models and not features we use the dataset in Duygulu et al.[9] . This also allows us to compare the performance of models in a strictly controlled manner. The dataset consists of 5,000 images from 50 Corel Stock Photo cds. Each cd includes 100 images on the same topic. Segmentation using normalized cuts followed by quantization ensures that there are 1-10 blobs for each image. Each image was also assigned 1-5 keywords. Overall there are 371 words and 500 blobs in the dataset. Details of the above process are contained in Duygulu et al [9]. We divided the dataset into 3 parts - with 4,000 training set images, 500 evaluation set images and 500 images in the test set. The evaluation set is used to find system parameters. After fixing the parameters, we merged the 4,000 training set and 500 evaluation set images to make a new training set. This corresponds to the training set of 4500 images and the test set of 500 images used by Duygulu et al [9].

VIII. EVALUATION OF AUTOMATIC IMAGE ANNOTATION

The FAM model uses a fixed number of words to annotate the images. To evaluate the annotation performance, we retrieve images using keywords from the vocabulary (note that this is not ranked retrieval). We can easily judge the relevance of the retrieved images by looking at the real (manual) annotations of the images. The recall is the number of correctly retrieved images divided by the number of relevant images in the test dataset. The precision is the number of correctly retrieved images divided by the number of retrieved images. We calculate the mean of precisions and recalls for a given query set. To combine recall and precision in a single efficiency measure, we use the F-measure,

$$F = \frac{2*recall*precision}{recall+precision}$$

TABLE 1: EVALUATION OF RANKED RETRIEVAL

Table 1 showsthe different query sets and relative performance of the two retrieval models in terms of average precision.Table 1 shows the details of the four subsets of our query set, along with average precision for the two retrieval models on each of the subsets. We achieve average precision of above 0.2 on 3-4 word queries. This is particularly encouraging because the results are obtained over a large number of queries. As expected, performance is generally higher for longer queries. The direct retrieval model (DRM) outperforms the annotation-based model (PAM) on all query subsets. The differences are statistically significant according to the Wilcoxon test at the 5% confidence level. The exception is the 4-word query set, where it is not possible to achieve statistical significance because of the small number of queries.

TABLE 2: RECALL/PRECISION OF “DAY” AND “NIGHT”

		Day	**Night**
Original Annotation	Recall	0.67	0.00
	Precision	0.47	0.00
Modified Annotation	Recall	0.5	0.58
	Precision	0.74	0.25

Table 2 shows Recall/precision of “Day” and “Night” keywords before and after correcting erroneous manual annotations.

IX. CONCLUSION

The Multimedia mining, knowledge extraction plays crucial role in multimedia knowledge discovery. The paper was discussed in the use of Multimedia database management systems and also mining of different kinds of Multimedia data. Process of Multimedia mining is also discussed. Some of the issues in Multimedia mining are too much of data is lost when the sequence of multimedia is ignored. But in audio and video mining a basic problem rises which is a combination of information across multiple media. For annotating and retrieving images, three different models were suggested and tested. The FAM model is more than twice as good in terms of mean precision as a state of the art Translation Model in annotating images.

REFERENCES

[1] Jiewei Han andMicheiveKamber “Data Mining : Concept and Techniques”,2nd Edition, MorganKanfman Publishers,2006 ,Chapter 10 “Mining Object, Spatial, Multimedia, Text, and Web Data”

[2] Katashi Nagao, Shigeki Ohira, Mitsuhiro Yoneoka “Annotation-based multimedia summarization and Translation”

[3] Pravin M. kamde, Dr. siddu, P. Algur, “A Survey onweb mining Multimedia”/The International Journal of Multimedia & Its Applications (IJMA).

[4] Wijesekera D. and D. Barbara, “Mining cinematicknowledge: Work in progress”, in Proc of International Workshop on Multimedia Data Mining (MDM/KDD’2000), Boston, pp. 98–103.

[5] http://en.wikipedia.org/wiki/Automatic_image_Annotation

[6] W. B. Croft. Combining Approaches to Information Retrieval, in Advances in Information Retrieval ed. W. B. Croft, Kluwer Academic Publishers, Boston, MA.

[7] http://ciir.cs.umass.edu/~manmatha/papers/sigir03.pdf

[8] http://users.dsic.upv.es/~rparedes/research/papers/pdfs/Villegas12_CERI_WEBUPV.pdf

[9] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In 7th European Conference on Computer Vision, pages 97-112, 2002.

[10]Berger,A and Laerty, J. Information retrieval as statistical translation. In Proceedings of the 22nd annual international ACM SIGIR conference, pages 2221999.

[11]D. Hiemstra Using Language Models for Information Retrieval. PhD dissertation, University of Twente, Enschede, The Netherlands, 2001.

[12]J. M. Ponte, and W. B. Croft, A language modeling approach to information retrieval. Proceedings of the 21st annual international ACM SIGIR Conference, pages 275–281, 1998.

[13]V. Lavrenko and W. Croft. Relevance-based language models. Proceedings of the 24th annual international ACM SIGIR conference, pages 120-127, 2001.

[14]V. Lavrenko, M. Choquette, and W. Croft. Cross-lingual relevance models. Proceedings of the 25th annual international ACM SIGIR conference, pages 175–182, 2002.

[15]J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval, Proceedings of the 24th annual international ACM SIGIR Conference, pages 111-119, 2001.