Multiple Topic Detection by Parametric Mixture Models (PMM)—Automatic Web Page Categorization for Browsing

Kazumi Saito†

Abstract

This article explains a statistical model called PMM (parametric mixture models) for efficiently detecting multiple topics in documents such as Web pages describing a wide variety of contents. It also shows that in experiments using a set of real Web pages, PMM outperformed conventional methods in terms of topic detection performance.

1. Introduction

What do you do when you want to find a specific fact such as the height of Mt. Fuji? The traditional way was to look in a reference book at home or in a library, but these days many people perform a Web search using keywords such as "height" and "Mt. Fuji". This can quickly retrieve the desired information from among the huge number of pages on the World Wide Web. On the other hand, we sometimes want to learn something new by browsing documents. In this article, I mainly focus on the latter situation.

In general, we can easily browse even a huge number of documents and books if each one is suitably categorized according to its topic. For instance, libraries employ a standard classification system for categorizing books to make it easy to find books. However, a general document may have more than one topic, so it may be difficult for a librarian to decide on a single place to put a book about visiting historical places in Kyoto. Should it go under travel, society, or history? It might be best to regard the book as having all three of these topics and put one copy in each of these locations.

Web pages let authors freely write about their own intentions, opinions, etc., so they often cover multiple topics. To cope with this situation, we need to categorize a document using multiple topics. However, when considering a wide variety of topics, the total number of combinations of multiple topics becomes extremely large. While it would be impossible to prepare physical shelves in a library to hold multiple copies of books, we can easily prepare multiple virtual shelves for multi-topic documents by making full use of hyperlinks.

2. Web page categorization by directory

Many portal sites are equipped with a directory service where Web pages are categorized into multiple topics by humans. For instance, the open directory project [1] constructed a hierarchical classification directory where its top level consists of 16 general topics (categories) and each of these topics is in turn decomposed into several specific subtopics (**Fig. 1**). Many pages can be found by following more than one path. For example, you can find the home page of NTT Communications by following a series of hyperlinks starting under either "Business" or "Regional" on the top-level page of the open directory. We can regard this Web page as a typical example of a multi-topic document, i.e., one that refers to both "business & regional" information.

It is difficult to place every Web page into a directory. Namely, since the number of Web pages is increasing every day, it is highly labor intensive for humans to assign a suitable multi-topic label to each Web page one by one. Therefore, it would be very convenient if a personal computer could automatically determine a suitable multi-topic label for each Web page. To this end, by considering the intrinsic nature of multi-topic documents, NTT Communication Sci-

† NTT Communication Science Laboratories
Soraku-gun, 619-0237 Japan
E-mail: saito@cslab.kecl.ntt.co.jp

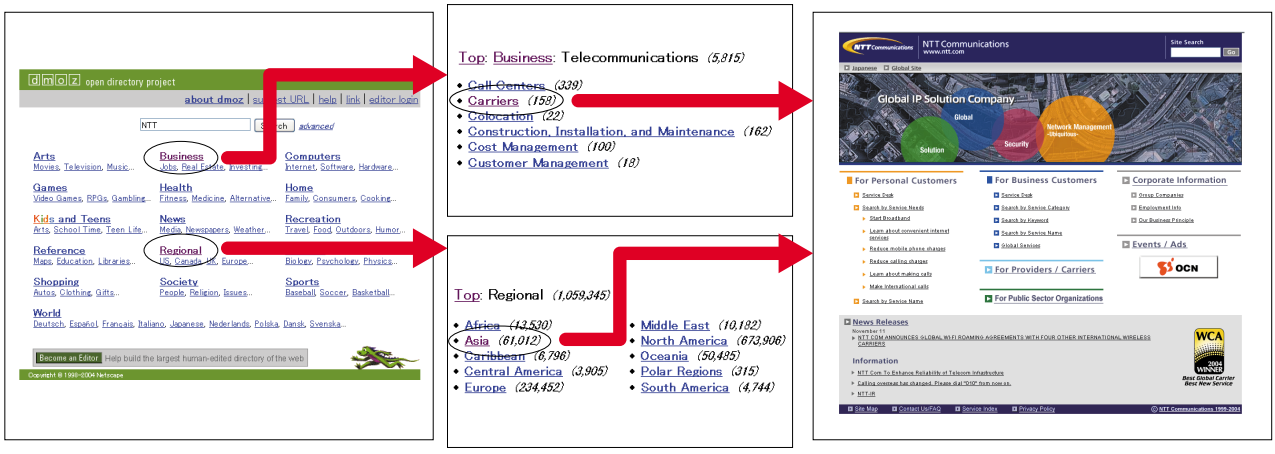

Fig. 1. Web page classification by directory in a portal site.

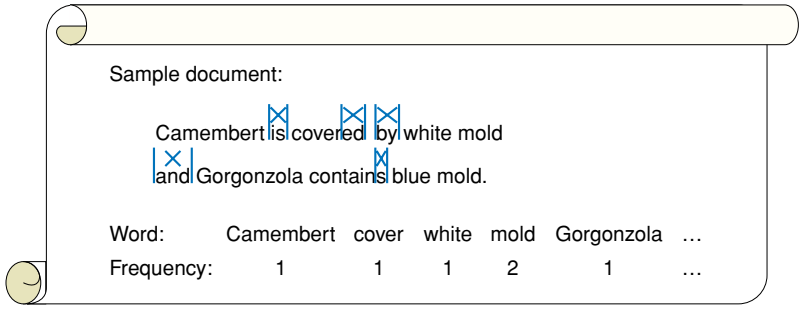

Fig. 2. Document representation by word frequency.

ence Laboratories has developed a multiple topic detection model called PMM (parametric mixture models).

3. Parametric mixture models

To produce an input data set for PMM, we first analyze a document (Web page) by decomposing it into a set of words and counting the frequency of each word. We exclude 571 well-known stop-words such as “is” and “by” and truncate some words such as conjugated verbs like “cover” (**Fig. 2**). Here, since the appearance order of words is completely ignored, the document representation based on word frequency is called a bag-of-words expression. This expression is widely used in the text mining field.

Now, consider the frequency of words appearing in multi-topic documents. By examining actual Web pages related to the topic “cheese”, we found that the characteristic word set of the Web pages whose dual-topic category is cheese “sales & recipes” consists of a mixture of the two sets of characteristic words: that for Web pages on the single topic “sales” and that for pages on “recipes” (**Fig. 3**). Based on this insight, my colleague and I developed PMM [2]. Namely, the main contribution of PMM is that this natural insight is mathematically formalized as a statistical model.

From a set of documents labeled by allowing multiple topics, we can automatically construct the PMM model by using a learning algorithm. Therefore, if the training data set is replaced, we can easily construct a model for another topic detection system. Since PMM consists of general-purpose techniques, we can easily apply it to a set of general documents, not just Web pages. Of course, PMM can also be used to detect an unknown multi-topic label of a new document.

There are many conventional methods for document categorization. However, the statistical methods such as naive Bayes (NB), as well as pattern recognition methods such as support vector machines (SVM), cannot simultaneously detect multiple topics of documents. Some existing methods, such as the k-nearest neighbors (k-NN) method and neural network (NN) method, can detect multiple topics in documents, but they do not consider the intrinsic proper-

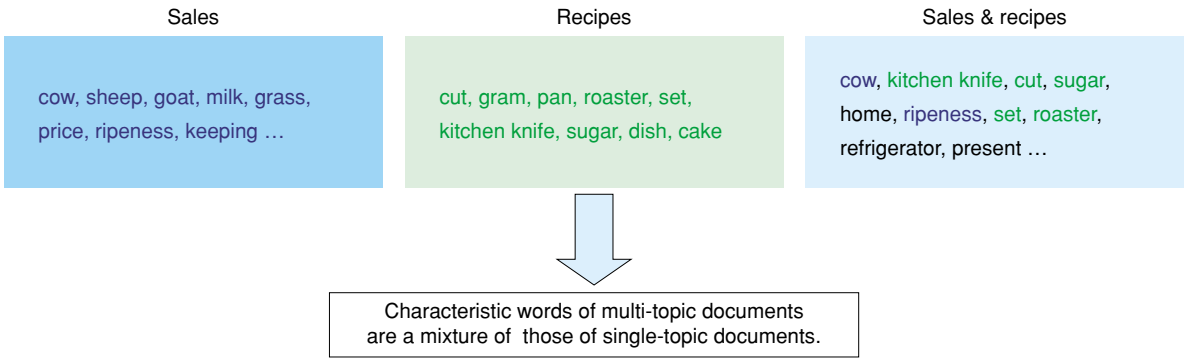

Fig. 3. Basic idea of parametric mixture models.

Table 1. Topic detection performance of PMM.

Problem name	Basic statistics			Multi-topic detection performance				
(top category)	Vocab.	Words	Topics	NB	SVM	kNN	NN	PMM
Arts & humanities	23146	111.1	26	41.6	47.1	40.0	43.3	50.6
Business & economy	21924	102.1	30	75.0	74.5	78.4	77.4	75.5
Computers & Internet	34096	128.2	33	56.5	56.2	51.1	53.8	61.0
Education	27534	111.8	33	39.3	47.8	42.9	44.1	51.3
Entertainment	32001	145.7	21	54.5	56.9	47.6	54.9	59.7
Health	30605	108.8	32	66.4	67.1	60.4	66.0	66.2
Recreation	30324	129.9	22	51.8	52.1	44.4	49.6	55.2
Reference	39679	163.7	33	52.6	55.4	53.3	55.0	61.1
Science	37187	173.3	40	42.4	49.2	43.9	45.8	51.4
Social science	52350	154.4	39	41.7	65.0	59.5	62.2	61.1
Society & culture	31802	176.2	27	47.2	51.4	46.4	50.5	54.2

ties of documents with multiple topics, so we think that they will have natural limitations for this task.

4. Application to document classification

To evaluate the topic detection performance of PMM, we performed experiments using a set of Web pages stored in the directory of the popular Japanese portal site “goo”, which is similar to Yahoo. We conducted separate experiments on each of the 11 top-level categories and regarded their second-level categories as multiple topics to be detected [3]. For each category, **Table 1** shows basic statistics, i.e., the vocabulary size (the total number of different words), the average number of words in one Web page, and the number of single topics.

For each of these 11 categories, Table 1 compares the topic detection performance of PMM with those of the main conventional methods: NB, SVM, k-NN, and NN. In the experiments, we used 2000 documents as a training set and a different set of 3000 documents as the test set. In the evaluation, we employed the F-measure, which is widely used in the field of information retrieval. The F-measure is 100% if all of the topics are perfectly detected and 0% if any one of the topics cannot be detected. In the table, the numbers in red show the best performance among the methods used in our experiments. As shown in the table, PMM outperformed the conventional methods by up to 10%, indicating its superiority over conventional methods. Using a 2.0-GHz Pentium PC, the training time for PMM was about 4 minutes (average for 11 problems) and the test time using 3000 pages was about 1 minute, indicating that PMM is very efficient especially compared with k-NN or NN.

5. Application to document retrieval

Although keyword search is widely used to retrieve desired documents, we cannot obtain documents that do not contain adequate keywords. It would be very convenient if we could retrieve documents by topic

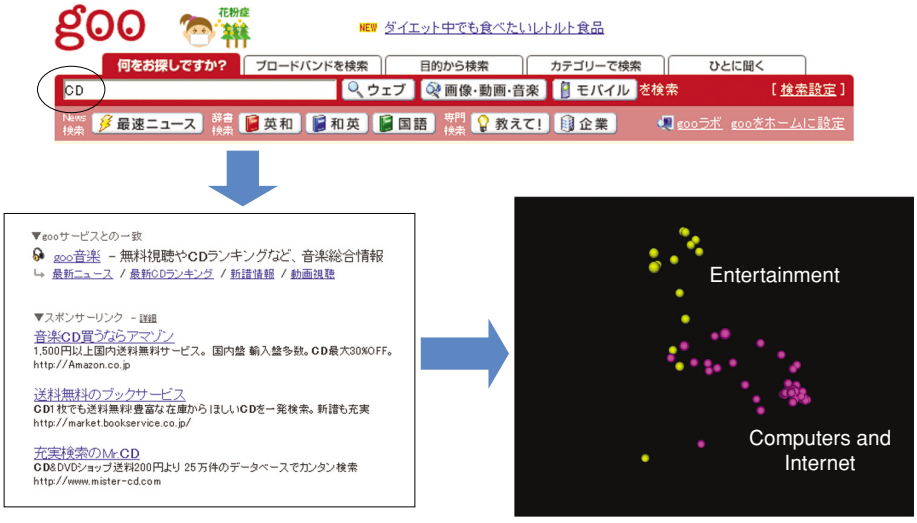

Fig. 4. Visualization of Web page retrieval results based on topic similarity.

similarity [4]. In our experiment using real Web pages, a Web page describing a study of the problems with the natural environment enabled us to retrieve some related Web pages describing environmental protection and solar batteries even though these pages did not explicitly contain any of the keywords "natural environment problems". This shows that PMM can retrieve Web pages that a standard keyword search fails to retrieve. Moreover, by replacing the set of documents used for training PMM, we can retrieve Web pages with different aspects.

We can also use PMM to arrange the retrieved Web pages based on topic similarity. **Figure 4** shows a three-dimensional visualization map obtained by keyword searching using a term "CD" in the site "goo". This visualization method is explained in detail in the next feature article in this issue[5]. Each Web page is represented by a colored circle, where one color indicates the topic "Entertainment" and the other "Computers & Internet". We can easily distinguish the two main fields for CD, i.e., compact disc contents (software) and compact disc equipment (hardware). Note that this experiment was conducted using Japanese Web pages.

6. Conclusion

PMM (parametric mixture models) is a new statistical model for efficiently detecting documents about multiple topics such as Web pages describing a wide variety of contents. Our experiments using a set of real Web pages showed that PMM outperformed conventional methods in terms of topic detection performance. We believe that PMM can be a basic model for constructing various Web services.

References

[1] http://dmoz.org/
[2] N. Ueda and K. Saito, "Parametric mixture models for multiple-topic text," Trans. of IEICE, Vol. J87-D-II, No. 3, pp. 872-883, 2004.
[3] N. Ueda and K. Saito, "Single-shot Detection of Multi-category of Text using Parametric Mixture Models," Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD2002), Edmonton, Alberta, Canada, pp. 626-631, 2002.
[4] N. Ueda and K. Saito, "Multiple-topic model for topic-based text retrieval," Trans. of IPSJ, Vol. 44, No. SIG14 (TOM 9), pp. 1-8, 2003.
[5] T. Yamada, "Visualizing Large-scale Network Data Structures," NTT Technical Review, Vol. 3, No. 3, pp. 19-23, 2005.

Kazumi Saito
Distinguished Technical Member, Senior Research Scientist, Emergent Learning Research Group, Intelligent Communication Laboratory, NTT Communication Science Laboratories.
He received the B.S. degree in mathematics from Keio University, Yokohama, Kanagawa in 1985 and the Ph.D. degree in electrical engineering from the University of Tokyo, Tokyo in 1998. In 1985, he joined the Electrical Communication Laboratories of NTT. He has received the Funai Best Paper Award of FIT2004, Best Paper Award of the Japanese Society of Artificial Intelligence, and Best Paper Award of the Information Processing Society of Japan. He is a member of IEEE, the Institute of Electronics, Information and Communication Engineers of Japan, Japanese Neural Networks Society, Japanese Society of Artificial Intelligence, and Information Processing Society of Japan.