CGM Data Mining Technology

MORINAGA Satoshi, YAMANISHI Kenji

Abstract

Due to the advancement of user participated web services, consumer generated media (CGM), such as grapevine communications on blogs and bulletin boards, are about to influence real society enormously. This paper introduces a number of text mining technologies intended for an integrated analysis of such CGM information (cyber space), television, newspapers and other such journalistic information (real space). It will also introduce the analysis corner of the "BIGLOBE Shunkan Ranking," in which analysis results are provided by the NEC Data Mining Technology Center, as an example of an actual analysis service that uses such technologies.

Keywords

CGM data, text mining, dynamic topic analysis, distributed cooperative topic analysis, contextual mining

1. Introduction

Due to the advancement of user participated web services, consumer generated media (CGM), such as grapevine communications on blogs and bulletin boards, are about to influence real society enormously. Conventional media, such as the television and the press influences on CGM; on the other hand, in some circumstances, topics delivered by CGM are picked up by conventional media. In such situations, it is important to understand the convergence of the information from cyberspace (such as CGM) and real space (such as media information), to analyze their correlation, as well as to see their transition in order to gain an overall view of the topics worldwide for the purpose of marketing analysis or the provision of new user participated web services.

A number of text mining technologies, used to integrate the analysis of CGMs and press information, are first introduced in this paper. The core of the matter is the possibility to conduct an integrated analysis on the (1) dynamic and (2) hetero data, along with (3) the ability to gain an overall view of the presented content. Dynamic topic analysis, distributed cooperative topic analysis, as well as key semantic mining, are introduced as mining methodologies that are able to respond to such a requirement.

This paper then introduces the "Analysis Corner" of the "BIGLOBE Shunkan Ranking," where the NEC Data Mining Technology Center is providing analysis results, as an actual case example of the integrated analysis service, which combines blogs, television broadcasts and Internet searches.

2. Dynamic Topic Analysis

Capturing topic structures and their changes from text streams that are delivered periodically over time, is a fundamental procedure for blog and press analysis. The framework of dynamic topic analysis[1], in item mentioned above, which detects the emergence of new topics by dynamically clustering a series of text strings, is herein introduced. The term "topic" used here represents a group of text strings describing a specific phenomenon or activity.

In the dynamic topic analysis the chronological data of text is used as input to execute the following tasks:

1) Topic structure identification: Discover the structure regarding what topic is found in what rate.

2) Topic emergence detection: Detects the emergence of a new topic and cessation of an existing topic in a timely manner.

3) Topic characterization extraction: Extracts expressions that are characteristically featured in individual topics.

The following issues have been formulated (**Fig. 1**) in order to make it possible to perform these tasks in a successive

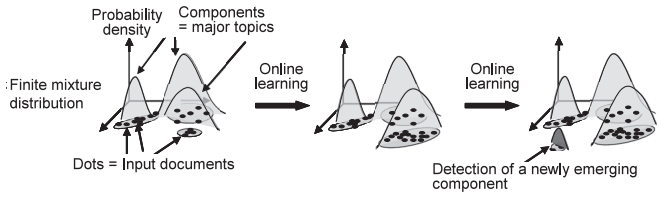

Fig. 1 Dynamic topic analysis.

manner:

(1) Modeling: First, a text string is expressed as a multi-dimensional vector that indicates the frequency of an individual word emergence or tf-idf values as elements of a text string. The statistical emergence structure for the topics is then modeled using the finite mixture model. The individual components (normal distribution or binary distribution), which constitute the finite mixture model, represents one topic and the mixture ratio represents the emergence probability distribution of topics.

(2) Learning: The aforementioned topic structure is identified by learning the topic structure online, using an online discounting EM algorithm with time stamp. The online discounting EM algorithm here, is an EM algorithm designed to learn even from transient data by gradually forgetting the influence of past data (for task 1).

(3) Determining optimum number of topics: The optimum number of topics, which transitions with time, is selected using a dynamic model selection. The dynamic model selection mentioned here is a function for obtaining the optimum number of mixtures for the finite mixture model, which changes with time in a dynamic manner. The emergence of new topics can also be detected through the detection of an increase in the number of mixtures (for task 2).

(4) Analyzing character of topics: Text strings that correspond to a mixture component are compared with those corresponding other components and characteristic words for each topic are derived through a ranking that is performed based on the measure of information, ESC (Extended Stochastic Complexity)[2] (for task 3).

Topic structures and their changes relating to continuously added data from CGMs and the press can be captured dynamically using the above mentioned framework.

3. Distributed Cooperative Topic Analysis

We consider an issue relating to deriving a topic structure representing the overall view by integrating information included in the data of text strings stored at scattered multiple remote sites. This is a basic issue for information integration. Here we aim to integrate information (1) from data that has hetero characteristics (2) without gathering raw data into one place (in order to protect privacy) and (3) with as few communications as possible (4) to achieve accuracy comparable with the results derived from collecting raw data into one place. The framework for distributed cooperative topic analysis[3] used to resolve this issue is shown below (**Fig. 2**).

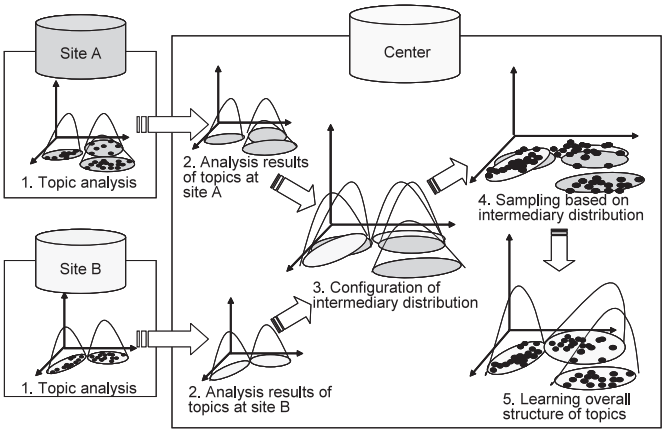

Fig. 2 Distributed cooperative topic analysis.

1) Topic structure at each individual site: A topic structure is identified for each site according to the methodology of the dynamic topic analysis described in Section 2. Only derived parameters are sent to the center.

2) Integration of information at center: Using dictionary knowledge to integrate hetero information, a mixture model, comprised by simply overlaying topics from individual sites, is created, based only on the parameters sent from each individual site. This model is called an intermediate distribution.

3) Overall structure through re-learning: Since similar topics are not summarized in the intermediate distribution, the overall structure is not yet visible. For this reason, re-learning of the topic structure is conducted by taking data again from the sampling based on the distribution of an intermediary product, which is then used to re-learn the topic structure. Let us call the derived overall model an overall topic structure.

Once the overall topic structure is made from the aforementioned framework, topics that are common at individual sites can be discovered and particular topics can be found at individual sites by clarifying the individual topics, as well as the relationship between individual topics with individual sites. This can be used to understand commonality and the distinctions between the topics from CGMs and the press.

4. Key Semantic Mining

The aforementioned individual technologies provide functions for "sorting out a large number of text string data to see whether or not they belong to the same topics." Key semantic

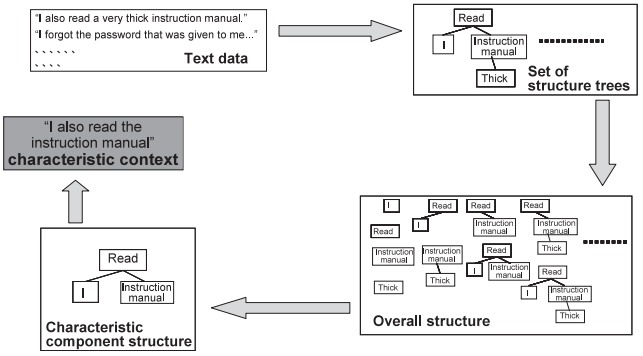

Fig. 3 Key semantics mining.

mining, on the other hand, is a technology for extracting and comprehensively viewing the characteristic content that appears in each individual group[7].

In key semantic mining the processes, such as (1) performing dependency analysis on sentence structure with components, for example, subject and qualifiers of basic sentence blocks in each sentence found in text strings, (2) extracting sentence structures that emerge more often in a group of particular text strings than others and (3) outputting generated Japanese language expressions that correspond to extracted sentence structures (**Fig. 3**).

Expressions output through such a process are indicated to make it possible for humans to understand "what is often described" and are used for understanding the context of CGMs or press reports on particular topics.

5. Integrated Analysis on CGMs and Press Reports by Shunkan Ranking

The "Analysis Corner"[4] of "BIGLOBE Shunkan Ranking" that prepares content based on analysis results, provided by NEC Data Mining Technology Center as an attempt to perform integrated analysis on cyber and real information, is introduced (refer also to "Analysis Blog"[5]). Here the purpose is to look into the world created by combining data derived from blogs, search results and television broadcasts from multiple angles using the aforementioned frameworks of dynamic topic analysis, distributed cooperative topic analysis and key semantic mining. So far special features on soccer's World Cup (July 2006), cinema during the summer break (August), games (September), onsen hot springs (October), dramas (November) and next-generation game machines (December), were presented.

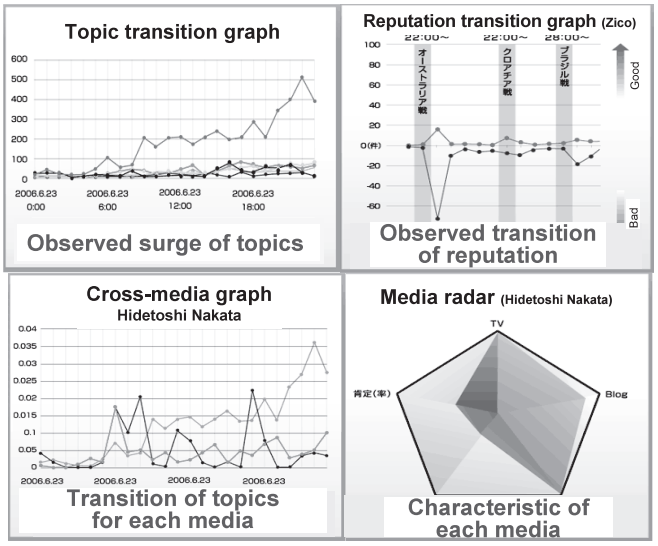

Fig. 4 Examples of Shunkan Ranking analysis.

The special feature on soccer's World Cup extracted reputation information[6] from blogs (provided by Datasection and NEC BIGLOBE), as well as data from television broadcasts (provided by Project) and displayed output, as shown in **Fig. 4**.

The topic transition graph chronologically indicates the kind of topics that are active as common topics for data from blogs and television broadcasts. The reputation transition graph indicates the chronological transition of favorable and unfavorable evaluations for particular athletes. The cross-media graph indicates chronological changes on the degree of exposure for particular athletes on blogs, television broadcasts and Internet searches. The media radar indicates a radar chart of particular athletes with regards to their appearance on television broadcasts, blogs, as well as ratios relating to opinions on them or search results, along with an abundance of opinions about them.

Furthermore, a graph depicting the amount of favorable opinions expressed on blogs (vertical axis) and the amount of time television advertisements were aired (horizontal axis) with ranking points representing the box office proceeds (size of bubbles), were depicted on a graph for movies covered by the special feature on cinema during the summer break, with an animated depiction of the changes on the graph since the beginning of their showing as time went on (**Fig. 5**). In this manner, it is possible to say that "bubbles flew upward with movies for which a lot of favorable opinions were expressed on blogs" or "those with bubbles flying off to the right had a lot of hours of television commercial coverage airing for the pur-

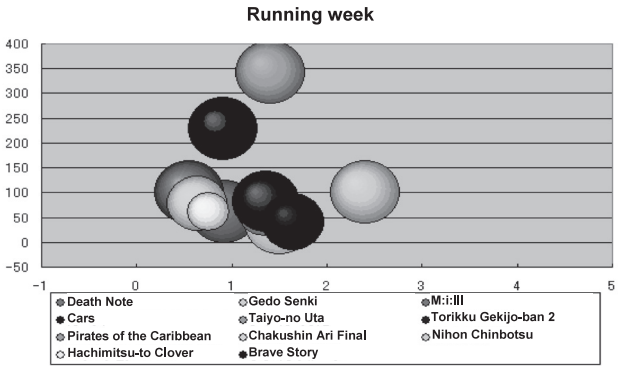

Fig. 5 Example of bubble chart.

First broadcast: Aired October 16, 2006
Rhapsody of a twisted pianist versus a self-centered conductor

1st	Identical to image portrayed by original novel
2nd	Starts today
3rd	Juri Ueno
4th	Naoto Takenaka
5th	CG
6th	Requiem
7th	Belly landing
8th	Hiroshi Tamaki
9th	"Pathetique"
10th	Tamagocchi

Second broadcast: Aired October 23, 2006
Drop out twisted orchestra! Having a rough start!?

1st	Test
2nd	Spring
3rd	Accompaniment with piano
4th	Falling over with foaming mouth
5th	Playing prank is fun
6th	Sonata for violin
7th	Enroll in another course
8th	Twisted orchestra
9th	Mine-kun
10th	Clumsy

Fig. 6 Example of characteristic blog entry.

pose of advertisement."

In the special feature for dramas, characteristic writings, in response to each airing of a particular television drama on blogs, were extracted (**Fig. 6**). It is possible to see the aspects of television broadcasts that have drawn the attention of the bloggers, who wrote about them on blogs.

Only a small portion of the analysis results was shown here. Refer to the corresponding pages and blogs for more details.

6. Conclusion

Examples of actual implementations are shown using text mining technologies for the analysis of CGMs and analysis of their conversion with press reports. By combining technologies, such as dynamic topic analysis, distributed cooperative topic analysis and key semantic mining, it was possible to capture a perspective that overviews surging topics in cyberspace, as well as their relationship with real space.

User participating web services are expected to develop further in the future and technologies for gaining an understanding of situations are expected to become important in the future.

* Some products and services introduced in this paper are mainly provided for the domestic market.

* The corporate and product names mentioned in this paper are trademarks or registered trademarks of their respective owners.

References

1) S. Morinaga and K. Yamanishi: "Tracking Dynamics of Topic Trends Using a Finite Mixture Model," in Proc. of KDD2004, ACM Press, 2004.
2) Yamanishi and H. Li: "Mining Open Answers in Questionnaire Data," IEEE Intelligent Systems. September/October, 2002.
3) Matsumura, S. Morinaga and K. Yamanishi: "BUNSAN HETERO NA Data KARANO Topic ZENTAIKOZO NO GAKUSHU (Learning overall structures of topics from distributed and hetero data)" (FIT2005).
4) http://search.biglobe.ne.jp/ranking/
5) http://mining.at.webry.info/
6) S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima: "Mining Product Reputations on the Web," in Proc. of KDD2002, ACM Press, 2002.
7) S. Morinaga, H. Arimura, T. Ikeda, Y. Sakao, and S. Akamine: "Kes Semantics Extraction by Dependency Tree Mining," in Proc. of KDD2005, ACM Press, 2005.

Authors' Profiles

MORINAGA Satoshi
Principal Researcher,
Common Platform Software Research Laboratories,
NEC Corporation

YAMANISHI Kenji
Research Fellow,
Common Platform Software Research Laboratories, and
Business Innovation Center,
NEC Corporation