Multifaceted Analysis of News Articles by Using Semantic Annotated Information

Masaharu Yoshioka
Hokkaido University
N14 W9, Kita-ku, Sapporo-shi,
Hokkaido, 060-0814, Japan
yoshioka@ist.hokudai.ac.jp

Noriko Kando
National Institute of Informatics
2-1-2, Hitotsubashi, Chiyoda-ku
Tokyo, 101-8430, Japan
kando@nii.ac.jp

ABSTRACT

We propose a novel framework that enables multifaceted analysis of news articles. This system uses semantic annotated information (e.g., person, place) as facets and can be used to construct structured queries for comparing the differences between sets of articles.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Query formulation; H.5.2 [**User Interfaces**]: Interaction styles

Keywords

Multifaceted analysis, Text mining

1. INTRODUCTION

Recently, we can access large amounts of texts from the Web. To understand the characteristics of these texts, it is better to have a mechanism for analyzing such texts from semantic points of views. EMM News Explorer [2] is one news analysis system with semantically annotated text. However, such systems do not have a good framework for comparing sets of news articles that satisfy different semantic conditions. For example, it is not easy to compare Japanese and American news articles about President Obama this month.

In this paper, we propose a multifaceted analysis framework that uses semantically annotated information (person, place, organization, news source, date, opinion, words) as facets. This framework can be used to construct structured queries for comparing differences between sets of articles.

2. USING SEMANTIC ANNOTATION FOR NEWS ARTICLE ANALYSIS

It is important for news article analysis to use semantic annotations, such as persons, places, and dates. Therefore, news analysis systems, such as EMM News Explorer[2], have a feature for adding such annotations to news articles and also have a feature for accessing news articles through such information. For example, a user can select the name of a person (e.g., Barack Obama) as a query and access news articles related to him. In addition, several semantically annotated pieces of information from selected news articles are aggregated in order to show related persons, places, and so on.

However, these systems do not have a good interface for constructing complex structured queries to analyze news articles for a specific topic. Therefore, it is not easy to analyze news articles with more specific conditions, e.g., key persons related to Barack Obama about "Global Warming".

For solving this problem, we propose a novel multifaceted interface for analyzing news articles that has the following features.

1. Selection of news articles by using a structured query with semantically annotated information
   We defined seven facets; person, place, organization, country of origin of a news source, date, opinion (positive or negative), and words, and they are usable to select articles for further anlysis. The user can specify a structured query by specifying the conditions with facet information.

2. Analysis of news articles
   The system enables characteristic information for each facet to be selected by using TF·IDF. By using this feature, the system can select related persons, places, and organizations for a given set of articles. The results can be displayed as a table or a bar chart with frequency information. In addition, the system can also use graphs to show how the number of articles that contain given facet information changes.

3. Comparison of analysis results by using different conditions
   To understand the information in the selected articles, it is better to compare the analysis results with similar cases. For example, when a user would like to find out key persons from a certain country, it is better to compare the related persons in a target country and ones in other countries.

4. Support analysis by encoding a specific analysis scenario as a template for the system
   Since it is not so easy for the users to set up such multiple facet analysis interfaces for comparison, the system has the capability of setting up windows for analysis by using templates, e.g., a comparison of related persons in different countries or a comparison of characteristic keywords between positive and negative articles.

ESAIR'12, November 2, 2012, Maui, Hawaii, USA.
ACM 978-1-4503-1717-7/12/11.

3. NSCONTRAST WITH MULTIFACETED ANALYSIS

The NSContrast system is a system for accessing news articles from multiple news sites [3]. This system has the pfollowing analytic modules.

- Term collocation analysis
  The system generates a list of characteristic terms by comparing news article databases from different countries. This term list is represented as a term collocation graph to aid in the understanding of the relationships among characteristic terms.

- A burst analysis function [1] for finding an appropriate time sequence window
  To find characteristic terms by using contrast set mining, it is preferable to select a large number of articles for a particular topic. Because burst analysis is a method that finds a period of time during which a given term is of more interest than usual, it is effective for finding this information.

- A news article retrieval system
  To understand the meaning of term collocation analysis results and burst analyses, a news article retrieval system is used.

This system supports semantic annotation by using Wikipedia[1] based ontology that are enhanced with GeoNames[2] data used for semantic annotation [4].

In addition to the previous analytic modules, we add the multifaceted analysis system discussed in the previous section as a new analytic module.

Figure 1 shows a screenshot of the multifaceted analysis system. This system consists of the following two components.

- Components for constructing a multifaceted query
  The areas on the right side of the screen are used to construct a multifaceted query. Users can select facets and input keywords for selecting news articles. In Figure 1, the user selected articles that contain "バラク・オバマ" (Barack Obama) as a person.

- Components for representing facet information
  A user can add multiple windows for comparing information with different queries. In each window, he/she specifies the following information to represent facet information.

  - Facet and representation style
    He/she can select a facet to show its information and representation style (e.g., time sequence graph, tables, and bar chart).
  - Additional conditions for making a structured query
    To allow for comparison between different conditions, the user can select additional queries for selecting news articles.

In Figure 1, the user added four windows that represent positive and negative articles. Each window has the country of origin of the news source (Japan, Korea, USA, and China) as an additional query. By using these graphs, he/she can analyze the trends of opinion in each country by comparing the data of different countries.

When the user modifies a query for selecting articles (e.g., by adding query keywords such as "global warming" and "economics"), all of these graphs are updated simultaneously. This is a novel feature that enables users to construct meaningful multifaceted queries for comparing differences. This is helpful to compare different trends of opinion about Barack Obama's various policies in each country.

Such layouts and additional queries are stored as analysis templates. The user can select one of the templates for a different analysis, e.g., a comparison of related terms for positive and negative articles.

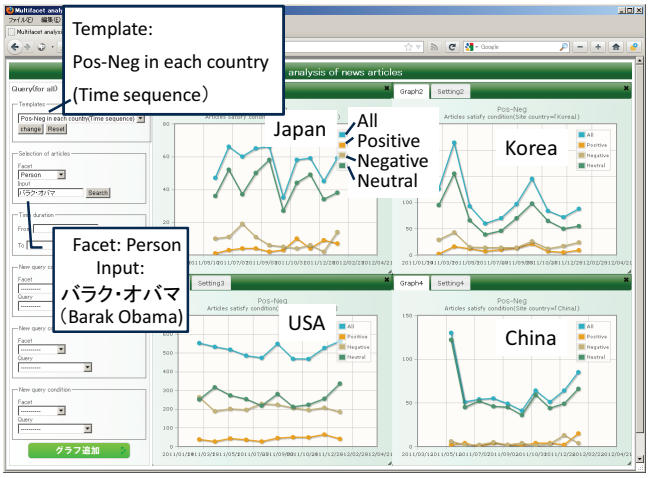

Figure 1: Screenshot of the multifaceted analysis system

4. SUMMARY

In this paper, we proposed a novel framework that enables multifaceted analysis of news articles. This interface is helpful for using semantically annotated data for detailed analysis.

5. REFERENCES

[1] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of the 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pages 91–101, New York, NY, USA, 2002. ACM Press.

[2] R. Steinberger, B. Pouliquen, and E. van der Goot. An introduction to the europe media monitor family of applications. In *SIGIR 2009 Workshop Proceedings, Information Access in a Multilingual World*, pages 1–8, 2009.

[3] M. Yoshioka. NSContrast: An exploratory news article analysis system that characterizes the differences between news sites. In *SIGIR 2009 Workshop Proceedings, Information Access in a Multilingual World*, pages 25–29, 2009.

[4] M. Yoshioka. ABRIR at NTCIR-9 geotime task usage of wikipedia and geonames for handling named entity information. In *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Quesiton Answering, And Cross-Lingual Information Access*, pages 75–81, 2011.

[1] http:// www.wikipedia.org/
[2] http://www.geonames.org/