Integration of Knowledge on Wikipedia and Other Web Resources

Eklou Damien† Yasuhito Asano‡ and Masatoshi Yoshikawa‡

† ‡ Graduate School of Informatics, Kyoto University Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, JAPAN

E-mail: † eklou@db.soc.i.kyoto-u.ac.jp, ‡ {asano,yoshikawa}@i.kyoto-u.ac.jp

Abstract:

Looking for desired information on the web can be a time consuming task. In this process Wikipedia constitutes a very helpful tool as it is the largest and most popular general reference site on the internet. Most search engines actually rank Wikipedia pages among the top listed results. However due to the nature of Wikipedia which is manually updated by users, it is virtually impossible to have all the valuable information related to a subject covered in a single article. In order to support the user search experience, we propose a method for finding valuable information not included in Wikipedia from other web resources.

Keyword: Complementary Information Retrieval, Wikipedia, Information Integration, Topic Modeling

1. Introduction

In other to constitute a base of organized knowledge available online we have seen the development of online encyclopedias recently. A representative of these efforts is the free online encyclopedia Wikipedia. Wikipedia distinguishes itself from its peers as it based on a collaborative approach which enables individuals from all over the world to add their contribution. Therefore articles in Wikipedia cover a broad range of domains that can’t be found in other encyclopedias. Also Wikipedia articles are often updated quickly and, as a result the coverage of current events is quite extensive. These advantages helped Wikipedia gain a large success among users making it the first reference site in the world and ranked seventh among all websites in the world [1].

However, due to the huge amount of information about a subject, it is virtually impossible for Wikipedia editors to cover all the important aspects related to a subject. Therefore we have a limitation of the information available to users in a sense that users can wonder if the particular Wikipedia page they are browsing contains all the information related to the subject they are interested in. Is there any valuable information not contained in this article but which can be found in other web pages? Manually searching for these information can be a time consuming process.

As an approach to solve this problem, we present a method for automatically retrieving valuable information not contained in the current article (here Wikipedia article) the user is browsing but related to the article’s subject and present them to the user. The basic idea of the proposed method is as follows:

Given a Wikipedia article on a certain subject that a user is browsing, we extract the topics from that article, and then we mine the web in other to find web pages related to that subject. We also proceed at a topic extraction from these retrieved web pages. Then, we compare the topic extracted from Wikipedia and those extracted from the other web pages in order to find new topics not included in the Wikipedia article.

Automatic retrieval of these complementary information would contribute greatly to the user search process by saving the time spent to manually analyze other web pages or articles in order to gain maximum information. Also, our method could be very useful in terms of improving Wikipedia as it can serve as an indicator in updating Wikipedia articles content.

2. Related Work

Liu et al. [2] proposed a system called WebCompare for web page comparison in order to find “unexpected information”. The user first specifies the (URLs) from its competitors’ Web sites, then WebCompare discovers the pages that contain unexpected information (for example services or products that its competitor offers…) with respect to the user’s Web site. The unexpectedness of a Web page is calculated using a measure based on representing each page of the web sites using the vector space model (tf/idf weighting scheme). This approach is different from ours as our primary target is the web and also we take into account the multiplicity of topics of documents.

In the area of cross-media, Q.Ma et al. [3] presented a topic content based method for retrieving news web pages related to a video sequence in order to provide the user with different viewpoints on the same topic. This approach is different from ours as we focus on web pages primarily. Also general web pages might have multiple topics in contrast to news web pages which usually contain a single topic.

In [4] Yeung et al. proposed a framework for assisting users enriching Wikipedia articles that are written in different languages. The proposed method is based on the following tasks, first identifying new information that is contained in the document in one language (source document) but not in the document in another language (target document) and then providing the best place for insertion of a translated version of the new information from the source document into the target document.

In order to achieve their goal, the authors proposed a method based on phrase as a unit for comparison while ours is based on topic comparison.

3. Basics Concepts and Method

In our approach, we define complementary information as valuable information related to the general subject but not contained in the current page (Wikipedia article) the user is browsing.

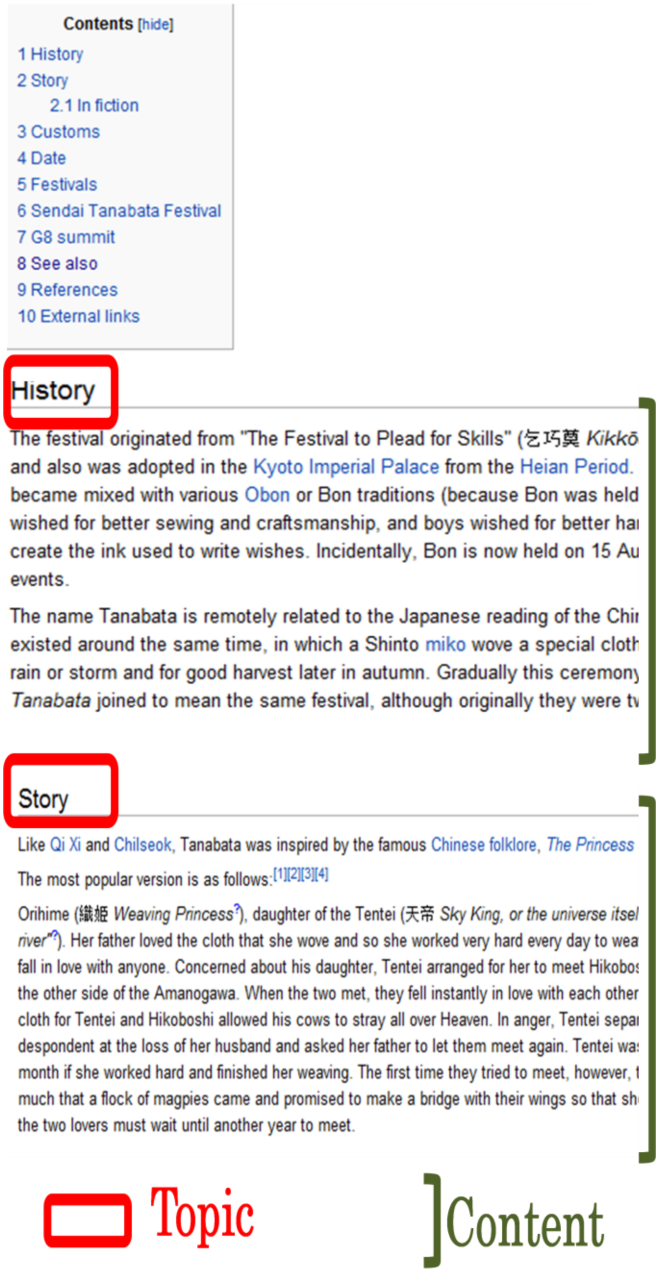
Our method is based on a topic comparison process. For a Wikipedia article on a certain subject, we extract the topics from that article; we query the web using the Wikipedia article’s title and perform the extraction of topics from the retrieved web pages. Finally we compare the Wikipedia topics and those extracted from the other web pages, this allows us to find topics not included in the Wikipedia article.

We model the topic contained in a Wikipedia article and in web pages retrieved from the web using the following two approaches:

- Topic–Content structure modeling for Wikipedia articles and
- Latent Dirichlet Allocation for web pages retrieved from the Web

3.1. Topic-Content structure :Using Wikipedia structure

We use the topic content structure for modeling the topics for a Wikipedia article. The Topic-content structure has been used in previous work such as in [4] in order to represent a document. Wikipedia having a well organized structure, we plan to take advantage of that structure thus having in most cases a well organized topic repartition of information related to a general subject (Figure 1).We define two kinds of keywords for our purpose: topic keywords and content keywords. Topics keywords are words appearing in a title or subtitle in a Wikipedia article. Content keywords are words appearing in the development of a topic. We represent a topic and its content by a vector containing the topic keywords and the content keywords.



Contents [hide]

1 History
2 Story
2.1 In fiction
3 Customs
4 Date
5 Festivals
6 Sendai Tanabata Festival
7 G8 summit
8 See also
9 References
10 External links

History

The festival originated from "The Festival to Plead for Skills" (乞巧奠 *Kikkō*
and also was adopted in the Kyoto Imperial Palace from the Heian Period.
became mixed with various Obon or Bon traditions (because Bon was held
wished for better sewing and craftsmanship, and boys wished for better ha
create the ink used to write wishes. Incidentally, Bon is now held on 15 Au
events.

The name Tanabata is remotely related to the Japanese reading of the Chi
existed around the same time, in which a Shinto miko wove a special cloth
rain or storm and for good harvest later in autumn. Gradually this ceremon
Tanabata joined to mean the same festival, although originally they were t

Story

Like Qi Xi and Chilseok, Tanabata was inspired by the famous Chinese folklore, *The Princess*

The most popular version is as follows:[1][2][3][4]

Orihime (織姫 *Weaving Princess*?), daughter of the Tentei (天帝 *Sky King, or the universe itsel*
river"?). Her father loved the cloth that she wove and so she worked very hard every day to wea
fall in love with anyone. Concerned about his daughter, Tentei arranged for her to meet Hikobo
the other side of the Amanogawa. When the two met, they fell instantly in love with each other
cloth for Tentei and Hikoboshi allowed his cows to stray all over Heaven. In anger, Tentei sepa
despondent at the loss of her husband and asked her father to let them meet again. Tentei wa
month if she worked hard and finished her weaving. The first time they tried to meet, however,
much that a flock of magpies came and promised to make a bridge with their wings so that sh
the two lovers must wait until another year to meet.

Figure 1: Wikipedia Topic-Content structure.

3.2. Topic retrieval from the web

One of the most successful methods for generative topic modeling is the Latent Dirichlet Allocation (LDA) proposed by Blei et al.[5]. LDA is a probabilistic generative model which assumes that every document is a distribution over a mixture of topics where a topic is a probability distribution over words.

We use Latent Dirichlet Allocation process which constitutes the state of the art generative model as a basis of our approach. The generative process associated with LDA is as follows:

- Sample K multinomial distributions (each of size V) φ_k from a Dirichlet distribution $Dir(\beta)$
- For each document
  (a) Choose distribution of topics $\theta_d \sim Dir(\alpha)$
  (b) For each word i in document d ($w_{i,d}$)
    i. Choose topic $z_{d,i} \sim \theta_d$
    ii. Choose word $w_{d,i} \sim \varphi_{z_{d,i}}$

Where α, β are Dirichlet parameters and V represents the vocabulary size.

We apply the LDA algorithm over the top M retrieved web pages with as query the Wikipedia article the user is browsing.

3.3. Complementary Information retrieval

The process of complementary information retrieval is based on the comparison of bag of words from the topics extracted from Wikipedia and those extracted from the Web by the LDA process. We compute the similarity (cosine similarity) between the two vectors. We suppose that topics vectors extracted by LDA which have the lowest similarities (under a certain threshold l) contain complementary information.

In order to eliminate topics not related to the subject from the set of potential useful topics, we proceed by the following approach:

(i) We compute the tf/idf for each word in the set of web pages and Wikipedia article.

(ii) We retrieve for each topic representing potential complementary information the top N web pages

(iii) We look for the occurrence of the top W (from (i)) words in the retrieved web pages.

The idea is to take the words most related to the Wikipedia article subject and check if they are contained in the web pages supposed to contain complementary information.

The retrieval of the pages containing the complementary information is done by retrieving web pages that contain a high proportion of the words related to the extracted new topic.

3.4. Presentation of the results to the user

We present to the user the retrieved topic related most significant words and also the web pages containing a high proportion of the words related to that topic.

Considering the fact that pages can contain very long textual content, we highlight the paragraph or area related to the topic to help users spot the targeted content quickly.

4. Future Work

We plan to finalize the design of our system. For the experiments we have to prepare the data set and do the preprocessing of these data. We plan to extract the textual contents of the different web pages in the data set. We plan to investigate how using the pages html tags could help us improve our method. We have to define the weighting scheme we will use for the topic and content terms.

In order to evaluate the complementarity of the extracted topics from Wikipedia and the other web resources, we have to define the metrics we plan to use.

We will also investigate ways to present the results to the users to help them quickly identify the related complementary information.

References

[1] Alexa, December 2010.

[2] Bing Liu, Yiming Ma, Philips Yu, “Discovering Unexpected Information from Your Competitors’ Web Sites.” KDD’01

[3] Qiang Ma, Katsumi Tanaka “Topic-Structure based Complementary Information Retrieval and Its Application” ACM Transaction on Asia Language Information Processing Vol 4, No 4,pp.475-503, ACM press 2005

[4] Ching-man Au Yeung, Kevin Duh and Masaaki Nagata “Assisting Wikipedia Users in Cross-Lingual

Editing”

WebDB Forum Tokyo, Japan November 2010

[5] Oyama S and Katsumi Tanaka

“Exploiting Document Structure for Comparing and Exploring Topics on the web“

WWW2003 (poster track)

[6] David M. Blei, Andrew Ng, Michael Jordan

“Latent Dirichlet Allocation”

Journal of Machine Learning Research 3(5), 993-1022 (2003)