IPSJ Transactions on Databases Vol.5 No.3 64–74 (Sep. 2012)

Regular Paper

How can the Web help Wikipedia? A Study of Information Complementation of Wikipedia by the Web

Damien Eklou[1,2,a)] Yasuhito Asano[1,b)] Masatoshi Yoshikawa[1,c)]

Received: March 20, 2012, Accepted: July 7, 2012

Abstract: With the huge amount of data on the Web, looking for desired information can be a time-consuming task. Wikipedia is a helpful tool because it is the largest, most-popular general reference site on the Internet. Most search engines rank Wikipedia pages among the top listed results. However, because many articles on Wikipedia are manually updated by users, several articles lack information and must be upgraded. That necessary information for updates can sometimes be found on the Web. Uprooting this information from the Web involves a time-consuming process of reading, analyzing and summarizing the information for the user. To support the user search process and to help Wikipedia contributors in the updating process of articles, we propose a method of finding valuable complementary information related to the Web. Experiments showed that our method was quite effective in retrieving important complementary information from Web pages.

Keywords: Web mining, Wikipedia, topic modeling

1. Introduction

Since the early days of the Internet, the amount of information available on the world wide web has been growing exponentially. Because the number of people that have the means to access the Internet is growing, the amount of data is also expected to grow even more rapidly as different technologies such as smart phones and tablets become increasingly common. Every person on the globe who has access to the Internet can upload contents, create blogs, have a diary, post information related to certain products, write reviews, and perform other tasks. Gleaning useful information from this humongous amount of data constitutes a great challenge of this era for information retrieval scientists.

One problem with information related to the Web is the fact that information is scattered through different sources such as Web pages and databases. For a user, looking for information about a certain topic going through all these information sources is expected to be time-consuming. Many studies have been conducted on that matter [4], [10].

To organize knowledge available on the Web, many online encyclopedias have been developed. One example of these efforts is the free online encyclopedia, Wikipedia. Wikipedia is based on a collaborative approach, i.e., Wikipedia enables individuals from all over the world to add their contributions to articles. This feature particularly helped Wikipedia gaining momentum between articles that it covers compared to other encyclopedias. In addition, one other quality of Wikipedia is the fact that Wikipedia articles are often updated quickly for certain domains of knowledge and the coverage of current events is extensive. These advantages helped Wikipedia gain large success among users, making it the first reference site in the world ranking among the top ten most visited websites in the world [*1].

One advantage of Wikipedia, as previously stated, is the fact that it contains a huge number of articles. However, one major problem among others with Wikipedia is that, many of these articles are still regarded as stub articles. In fact, a Wikipedia article [*2] reports that more than 35% of Wikipedia articles were stub articles in 2006. Although this number has probably improved over time, the existence of many stub articles cannot be denied. Wikipedia defines a stub article as an article having only few sentences of text which, although providing some useful information, is too short to provide encyclopedic coverage of a subject, and can be expanded. Wikipedia, in an effort to recognize these articles, has introduced a rating system for articles. Users are asked to give their feedback on the trustworthiness, objectivity, completeness, and whether the articles are well-written or not. Aside of articles marked as stub articles, there are many articles that do not offer good coverage of the subject they purportedly explain. On the other hand, such lacking information can sometimes be found on the Web. For example in the case of Yutaka Taniyama, who is a famous Japanese mathematician who committed suicide, the Wikipedia article provides no information about his background or hobbies. However information about these aspects of his life can be found on the Web. As in this example, one question that comes to the mind for most Wikipedia users is whether any valuable information not included in the Wikipedia article can be

[1] Graduate School of Informatics, Kyoto University, Kyoto 606–8501, Japan
[2] Division of Global Market Equity Technology, Deutsche Securities Inc. Japan, Chiyoda, Tokyo 100–6171, Japan
[a)] damience84@yahoo.co.jp
[b)] asano@i.kyoto-u.ac.jp
[c)] yoshikawa@i.kyoto-u.ac.jp

[*1] http://www.alexa.com
[*2] http://en.wikipedia.org/wiki/User:Dantheox

found on the Web.

One challenge with Web pages is that, in contrast to Wikipedia, they are usually not well organized with the information grouped into sections and subsections. Therefore manually searching for information can be a time-consuming process because the user must analyze each page to obtain the desired information or knowledge.

As an approach to improve these Wikipedia articles and to answer user needs for obtaining complementary information, we present a method of automatically retrieving valuable information from the Web and presenting it to the user. Our idea is based on topical analysis of Wikipedia articles and Web pages retrieved from the Web. Given a Wikipedia article on a certain subject, we use the Wikipedia article title as a query and use a search engine to retrieve Web pages. We proceed to conduct topic retrieval from the collection composed by the Wikipedia article and the retrieved Web pages. Subsequently, we process these topics to extract valuable knowledge. Automatic retrieval of valuable information from the Web is expected to contribute greatly to the user search process by saving time spent to analyze other Web pages or articles manually to gain maximum information. In addition, our method can be useful to improve Wikipedia because it can serve as an indicator to the users when updating Wikipedia article contents.

We conducted some experiments using Web pages that had been obtained by querying search engines. The results of our experiments showed that our method is useful to retrieve complementary information from the Web pages.

We modify the proposed method in the preliminary version [7], and reexamine all the experiments described there. In addition, we newly conduct experiments for comparing our method with a naïve method utilizing tf-idf.

2. Problem Formulation

Our goal is to gather valuable information from the Web to complement a Wikipedia article. A user can receive many pages related to the subject of a Wikipedia article using a search engine on the Web. This user, to obtain a full view of the subject, is obliged to analyze information included in the article and all pages obtained using the search engine. The user's search experience could be greatly improved if we were able to complement the article automatically using the information.

We target complementary information of two kinds:

Information improving topic coverage: information derived from topics that are not covered in the Wikipedia article.

Detailed information: information related to topics already covered but adding new specificities to the Wikipedia article content.

Both the information improving topic coverage and the detailed information are useful for the user search experience. The main problem here lies in how to model the information included in the article and all the pages obtained using the search engine considering the fact that a page and a Wikipedia article can address multiple topics. For a page covering multiple topics, it is challenging to find its part which contains complementary information about a certain topic. We use a topic model explained in Section 4.2.

We plan to retrieve the information from a topical perspective. We distinguish the following topics of two types to evaluate our proposal:

Human based topics: information that is linked thematically and which can be grouped into the same category by a human assessor.

Latent topics: topics obtained using a generative topic modeling process.

3. Related Work

Liu et al. [8] proposed a system called WebCompare for Web site comparison. The goal of this system is to extract information defined by the authors as "Unexpected Information," that is, information relevant but unknown to the user or contradicting the user's beliefs or expectations. For given competitive Web sites, their system gathers all Web pages. They represent each page as a vector using the term frequency inverse document frequency (tf-idf) weighting scheme and cluster them using the similarity of these Web page vectors.

Their goal is similar to ours in the sense that they attempt to retrieve relevant information that is unknown to the user. This unexpected information can also be regarded as complementary information because it complements knowledge related to the user about different aspects of these Web pages. The main differences of this approach from our method is the fact that our method's main target for information gathering is the Web and that our method considers the multiplicity of topics that can be addressed in the contents of a single Web page.

In the area of cross-media information retrieval, Ma et al. [10] presented a topic content based method for retrieving news Web pages related to a specified video sequence. Their goal is for a video sequence to provide the user with different viewpoints on the same topic. The process of information gathering is as follows: from a video sequence, they extract the caption contents and then generate subject-content based expanded queries to query general search engines to retrieve news articles related to the video. They defined the subject as the dominant terms in a video sequence and the content as the terms that have a strong co-occurrence relation with the subject terms. The main differences from our approach are that our emphasis is not solely on news articles but also on general Web pages. Furthermore, we incorporate consideration of the multiplicity of topics that can be covered in a Web page because general Web pages might have multiple topics, in contrast to news articles, which usually address a single topic.

Yeung et al. [17] proposed a framework for assisting users enriching Wikipedia articles written in different languages. The proposed method first takes two Wikipedia articles about the same subject which are written in different languages. The method then segments the articles using sentences as a unit. Each sentence is translated using machine translation services. They are mutually compared using vector representation for each sentence. Using this process, the proposed method identifies new information that is included in the document in one language (source document), but not in the same document written in another language (target document). Then it provides the best place for insertion of

a translated version of the new information obtained from the source document into the target document by label propagation techniques. To achieve their goal, the authors proposed a method based on phrases as a unit base for comparison. Their goal closely resembles ours because the target is to find information that is lacking in a Wikipedia article, but from a multilingual perspective. Our method differs from this approach because our method is based on a topic-level comparison and our goal is to complement the Wikipedia articles from information retrieved from the Web.

Nadamoto et al. [11] proposed a method for finding content holes, which is information that users are unaware of, in community-type contents of resources such as SNS and blogs. Their method relies on using Wikipedia as a source to obtain more general information about a community-type discussion subject. They compare threads in a community with articles on Wikipedia that cover the same topic and then extract coverage based content hole from Wikipedia. Their method is similar to ours because increasing the coverage is a target of this study, although their goal, which is upgrading the coverage of Wikipedia articles, is fundamentally different from ours.

Many studies have been done to obtain similar Web pages such as Refs. [4], [12], [13]. Nakatani et al. [14] proposed a method to rank the results obtained using a search engine according to the degree of difficulty with the easiest to understand being first. These reports describe various methods that are useful to obtain information from a certain perspective. However we are not aware of a method using a probabilistic topic approach for this purpose.

4. Our Approach

In this section, we present a method for retrieving information that is complementary to a Wikipedia article specified by a user. Our method can be summarized in the following five steps:

(1) Data collection: gathering Web pages related to the specified article; our data collection consists of the gathered pages and the article.
(2) Topic modeling: modeling data collection obtained in the previous step to find latent topics in the collection.
(3) Topic analysis: analyzing latent topics to find those that can complement the specified article.
(4) Sentences Extraction: Extracting the sentences that represent the latent topics found in the previous step.
(5) Presentation of the Results: Presenting the extracted sentences, combined with the article, to the users.

Step (1) is discussed in Section 4.1. Step (2), topic modeling of the data gathered in the previous step is conducted using Latent Dirichlet Allocation. Specific aspects of this step are explained in Section 4.2. The topic analysis Step (3) is described in Section 4.3. The sentence extraction process Step (4) constitutes an important part of our method because our goal is to extract the parts of Web pages that include the related complementary information; this step is explained in Section 4.4. Section 4.5 introduces ideas related to the presentation of the results to the user in a manner that is easy to access and navigate through. This constitutes the last step of our method.

4.1 Data Collection

Because our goal is to complement the Wikipedia article from the Web, the quality of the information retrieved from the Web is expected to have a direct impact on the information retrieved using our method. In the Web, we encounter disambiguation problems when dealing with results obtained from search engines. Because Step (1) of our method gathers pages using a search engine, it is important for certain Wikipedia subject to retrieve pages that are related to that subject and to discard those that are not. In this matter, we use the first sentence of the specified article. We take advantage of the writing style of most Wikipedia articles: the first sentence of an article is usually a general introduction of the definition of the article's subject. We presume that the words in this first sentence can be important identifiers to eliminate unrelated pages. From this sentence, we retain only the nouns. They constitute specific words that we use to test the relatedness of the Web pages obtained by querying the search engines. After selecting the specific words, we discard the Web pages obtained from the search engines based on whether they include those specific words or not. Web pages that include fewer than two of these words were discarded. We use Google search engine throughout our experiments.

The gathered pages and the specified article constitute our data collection. We proceed to the next step, in which we model data to extract the latent topics in data collection.

4.2 Topic Modeling Using Latent Dirichlet Allocation

Latent Dirichlet Allocation, abbreviated to LDA, is a probabilistic generative model proposed by Blei et al. [1]. LDA assumes that every document is a distribution over a mixture of topics for which a topic is a probability distribution over words, as depicted in **Fig. 1**.

Let K be the positive integer parameter specified as the number of latent topics, and let N be the size of the vocabulary. The generative process associated with LDA is the following:

- For each latent topic z_i for $1 \leq i \leq K$, sample a multinomial distribution φ_i of N words from a Dirichlet distribution $Dir(\beta)$ with Dirichlet parameter β.
- For each document d, do the following (a) and (b).

(a) Choose "topic distribution" θ_d of document d such that $\theta_d \sim Dir(\alpha)$, where α is a Dirichlet parameter.
(b) For each word n in document d, do the following (i) and (ii).

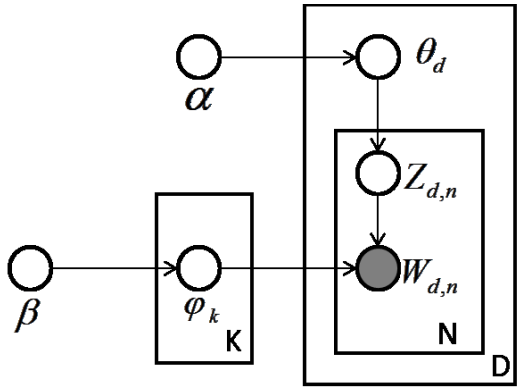

Fig. 1 LDA: a generative graphical model. D denotes the number of documents.

(i) Choose a topic $z_{d,n} \sim \theta_d$.
(ii) Choose a word $w_{d,n} \sim \varphi_{z_{d,n}}$.

Note that topic distribution θ_d of document d is the distribution of the probability of document d to be generated by each latent topic. For a set of documents, we can compute the topic distribution of the set by summing up θ_d of each document in the set and normalizing the sum.

The idea behind LDA is linked to human processes that take place while writing about a certain topic. First, a person has an idea of which topic about which we want to talk or write. Subsequently, that person chooses the words that can express the idea of that topic.

We regard LDA as suitable for our goal, which is to extract information from a topical perspective, because general articles on the Web can cover multiple topics. By applying LDA to a set of documents, we can obtain a topic distribution of an arbitrary subset of documents for K latent topics. Because the latent topics are common to the set of documents, we can compute the difference between the topic distribution of a subset of documents and that of another subset. In addition, LDA assigns values to words indicating their relevance to the different topics. These values can help in the sentence retrieval process using these values to assign each sentence a weight representing the relevance to each topic.

Let N be the number of Web pages retrieved by Step (1), as described in Section 4.1, and let a_w represent the specified Wikipedia article. Let $C = \{d_1, \ldots, d_N, a_w\}$ be the data collection obtained through Step (1). We apply LDA to the collection C to obtain the latent topics of C and to assess the probability that each document of $C = \{d_1, \ldots, d_N, a_w\}$ is generated by each latent topic. As explained above, applying LDA to C, we can obtain the topic distribution of an arbitrary subset of C for K latent topics common to all the documents. Therefore, by including a_w in C, we can obtain the topic distribution of C and that of a_w, called the topic distribution of the Wikipedia article, for the same K latent topics. Consequently, we can compute the difference of the topic distributions. If we apply LDA to $C' = \{d_1, \ldots, d_N\}$ and a_w independently, then the latent topics for C' should be different from those for a_w. Therefore, we could not compute the difference of topic distributions if we do not include a_w in C. We then analyze the difference to select which latent topics are not included in the Wikipedia article or which have a high probability of providing more detailed information than that which the article includes. This latent topic selection process is explained in the next section.

4.3 Topic Analysis

We compare the topic distribution of the Wikipedia article to the topic distribution of the entire collection C. Recall that the topic distribution of a document is the distribution of the probability of the document to be generated by each latent topic. Let $p(z_i \mid d_j)$ represent the probability that latent topic z_i generates document d_j. Let D be the set of documents that have a probability of being generated by latent topic z_i. Then the topic distribution of C is represented by the following formula:

$$p(z_i, C) = \frac{\sum_j p(z_i \mid d_j)}{|D|}.$$

Similarly, let $p(z_i, a_w)$ represent the topic distribution for a_w, i.e., the probability that topic z_i generates the Wikipedia article a_w.

When the probability value for latent topic z_i and the collection C is higher than the probability value for the same topic and the Wikipedia article a_w, i.e., when $p(z_i, C) - p(z_i, a_w) > 0$, we consider that a high probability exists of finding complementary information related to that particular latent topic in the collection. Therefore, this latent topic is selected in our approach. If the value $p(z_i, a_w)$ is near to zero for selected latent topic z_i, then the topic might not be covered in the Wikipedia article. In this case, information obtained from the Web for the topic would be classified into the information improving topic coverage explained in Section 2. Otherwise, the selected topic might be covered in the Wikipedia article partly. Then, information obtained from the Web for the topic would be classified into the detailed information explained in Section 2. A possible future work is to construct an automatic classification of the information obtained from the Web for each selected topic.

This analysis leads us to the next process that is extraction of sentences most related to the topic having a higher topic distribution in the collection than in the Wikipedia article.

4.4 Sentences Extraction

After comparison of the topic probability distribution of both the Wikipedia article and the collection, we proceed to sentence extraction from the selected latent topics that we regard as having a high percentage of complementary information with respect to the Wikipedia article. Let T represent the set of all the selected latent topics. The process retrieves the sentences most related to each topic $z_i \in T$. First, we select document d_{max} that has the highest probability of generating the topic z_i using the following formula:

$$d_{max} = \arg\max_{d_j}(p(z_i \mid d_j))$$

Then, we retrieve s sentences that have the highest score of representing topic z_i from the document d_{max}, where s is a positive integer parameter specified by the user. Let $w(x \mid z_i)$ represent the weight that each sentence x in the document d_{max} belongs to the topic z_i. In addition, let $p(w_r \mid z_i)$ be the probability that word w_r belongs to topic z_i, which is obtained by the LDA process. $w(x \mid z_i)$ is computed using the following expression:

$$w(x \mid z_i) = \sum_{w_r \in x} p(w_r \mid z_i),$$

We select the top s sentences according to the weight from the document d_{max} for each topic z_i.

4.5 Presentation of Results

The presentation of the results to users is an important issue because flooding the user with information is expected to produce the opposite of the desired effect. We plan to present the sentences most related to the topics selected by our method to the user. We also provide the link of the pages from which the sentences were extracted to enable the user to transition easily to the source article of the complementary information.

5. Experiments

In this section, we present our experimental settings and evaluation results. In Section 5.1, we present the data we use for our experiments. In Section 5.2, we define the metrics. Evaluation of the results is covered in Section 5.3. We present a running example in Section 5.4. Then, we conduct an experiment comparing the results of our method with those of a naïve method utilizing tf-idf and cosine similarity in Section 5.5.

5.1 Data

For each query, we gather the top 50 Web pages obtained by querying Google search engine. As a query, we use the Wikipedia article title. We conducted experiments with the following ten queries: "Yutaka Taniyama," "Influvac," "Izanagi," "Gozan no okuribi," "Soumaoro Kante," "Dufuna Canoe," "Comoe National Park," "Plumbosolvency," "Agflation," and "Bakanae." The words we use as queries and their definitions [*3] are presented in **Table 1**.

For each query, we process the Wikipedia article first sentences to extract the specific words for our page-discriminative approach explained in Section 4.1. These specific words are used to discard pages that are unrelated to the Wikipedia article subject. We used the Stanford NLP parser [*4] for this purpose. Specific words extracted from the first sentences are summarized in **Table 2**.

We used 10 pages for each query as the result of our page-discriminative approach. We used such a few pages for the evaluation to be effective because human evaluators reading many pages and extracting the topics present a difficult task. We processed the vocabulary by removing stop words and split each Web page at the sentence level for the sentence retrieval step.

We also evaluated the effectiveness of our approach to retrieve pages that are related to the Wikipedia article main subject. The goal is to determine if this approach is truly effective to discriminate irrelevant pages. We obtained a recall value more than 90% for our method for most queries. However, for Izanagi we obtained a recall value of about 18%, which results from the fact that mostly in the Web pages synonym of the specific terms retrieved by our approach were mostly used. Therefore, the use of synonyms can help improve the effectiveness of this approach.

Table 1 Queries.

ID	Queries	Definition
1	Yutaka Taniyama	Yutaka Taniyama (Japanese: 谷山 豊 Taniyama Yutaka; November 12, 1927, Kisai near Tokyo – November 17, 1958, Tokyo) was a Japanese mathematician known for the Taniyama–Shimura conjecture.
2	Influvac	Influvac is a sub-unit vaccine produced and marketed by Abbott Laboratories.
3	Gozan no Okuribi	Gozan no Okuribi (五山の送り火), more commonly known as Daimonji (大文字), is a festival in Kyoto, Japan.
4	Izanagi	Izanagi (イザナギ, recorded in the Kojiki as 伊邪那岐 and in the Nihon Shoki as 伊弉諾) is a deity born of the seven divine generations in Japanese mythology and Shinto, and is also referred to in the roughly translated Kojiki as "male-who-invites" or Izanagi-no-mikoto.
5	Soumaoro	Soumaoro Kante (var.: Sumanguru Kante) was a thirteenth century king of the Sosso people.
6	Dufuna Canoe	Dufuna canoe is an 8000-year-old canoe discovered by Fulani herdsman in Nigeria in 1987.
7	Comoe National Park	Comoe National Park is a national park in northeastern Cote d'Ivoire as well as a UNESCO World Heritage Site since its inception in 1983.
8	Plumbosolvency	Plumbosolvency is the ability of a solvent, notably water, to dissolve lead.
9	Agflation	Agflation, a term coined in the late first decade of the 21st century, describes generalized inflation led by rises in agricultural commodity prices.
10	Bakanae	Bakanae (pronounced "ba-ka-na-eh," not "ba-ka-nay.") or bakanae disease (Bakanae-byo), from the Japanese for "foolish seedling," is a disease that infects rice plants.

[*3] The definitions were extracted from the Wikipedia article about the same subject. We use the first sentence of the article in each case.

5.2 Metrics

To evaluate our complementary information retrieval process, metrics that can capture the effectiveness of our method constitute a key issue. The main reason is that it is not easy to quantify information. The purpose of our evaluation can be resumed in the following tasks:

- Is our method able to retrieve novel information from the set of Web pages obtained from the Web?
- To what extent are the retrieved sentences for a topic to be analyzed using our method semantically related?
- What is our method of coverage of the topics not included in the Wikipedia article, but which are covered in the Web pages?

We below explain the used metrics. In the following expressions, s represents the number of sentences retrieved for each latent topic. $|T|$ represents the number of latent topics selected using our complementary information retrieval process, as described in Section 4.4.

Table 2 Specific words.

Queries	Specific Words
Yutaka Taniyama	Japanese, November, Kisai, Tokyo, Mathematician, Conjecture, Taniyama-Shimura
Influvac	Vaccine, Abbott, Laboratories
Gozan no Okuribi	Daimonji, Festival, Kyoto, Japan
Izanagi	Kojiki, Nihon, Shoki, Deity, Divine, male-who-invites
Soumaoro	Sumanguru, Century, King, Sosso, People
Dufuna Canoe	Canoe, Year, Fulani, Herdsman, Nigeria
Comoe National Park	National, Park, Cote d'Ivoire, UNESCO, World, Heritage, Site, Inscription
Plumbosolvency	Solvent, Water, Lead
Agflation	Term, Decade, Century, Inflation, Rises, Agricultural, Commodity, Prices
Bakanae	Disease, Bakanae-byo, Seedling, Rice, Plant

[*4] http://nlp.stanford.edu/software/tagger.shtml

Purity

The s sentences for each selected latent topic can be regarded as a cluster. The purity metric helps evaluate the cohesion of the clusters. That is, we analyze the sentences retrieved for each selected latent topic and check whether the sentences are related to a common human-based topic. Let $CL(z_i, s)$ be the set of s sentences for selected latent topic z_i. Thus, $|CL(z_i, s)| = s$ for every z_i. Let $ht(z_i)$ be the number of human-based topics found in $CL(z_i, s)$. Then, $purity(z_i)@s$ representing the purity of the cluster $CL(z_i, s)$ of using s as a parameter is defined as

$$purity(z_i)@s = \max_{1 \le j \le ht(z_i)} m_j/s,$$

where m_j is the number of sentences judged as belonging to the j-th human-based topic. The overall purity using s as a parameter, denoted as $purity@s$, is the weighted average of $purity(z_i)@s$ for $1 \le i \le |T|$:

$$\begin{aligned} purity@s &= \sum_{1 \le z_i \le |T|} |CL(z_i, s)| purity(z_i)@s \Big/ \sum_{1 \le z_i \le |T|} |CL(z_i, s)| \\ &= \frac{1}{|T|} \sum_{1 \le z_i \le |T|} purity(z_i)@s. \end{aligned}$$

A high value of purity signifies that a high number of retrieved sentences for each selected latent topic are related thematically.

Novelty

This metric helps evaluate the degree to which new information is retrieved by our method for the selected latent topics. A high value of novelty implies that a high value of novel information is retrieved.

Let $N(z_i, s)$ be the set of sentences providing a new information among $CL(z_i, s)$, the top s sentences for selected latent topic z_i. Then, we define the novelty, denoted as $novelty@s$, using the parameter s mathematically using the following expression:

$$\begin{aligned} novelty@s &= \frac{1}{|T|} \sum_{1 \le z_i \le |T|} \frac{|N(z_i, s)|}{|CL(z_i, s)|} \\ &= \frac{1}{|T| \cdot s} \sum_{z_i} |N(z_i, s)|. \end{aligned}$$

Topic Coverage

This metric helps evaluate the extent to which human-based topics that are represented in the Web pages but which are not included in the Wikipedia article are being retrieved. Let $t_{collection}$ be the set of human-based topics found in the collection, and let $t_{Wikipedia}$ be the set of human-based topics found in the Wikipedia article. Let $t_{method}(s)$ be the set of human-based topics which belong to $t_{collection} \setminus t_{Wikipedia}$ and found in union $\cup_{1 \le i \le |T|} CL(z_i, s)$; the union is exactly the set of sentences retrieved for all the selected latent topics. Then, the topic coverage, denoted as $coverage@s$, using paramater s is defined as

$$coverage@s = \frac{|t_{method}(s)|}{|t_{collection} \setminus t_{Wikipedia}|}.$$

Because no well-known automatic approach exists for our purpose to evaluate the information retrieved by our method, we ask human evaluators to analyze the topic retrieved using our method and evaluate them.

5.3 Experimental Evaluation and Results

We describe the evaluation methodology and present the results we obtained for our dataset.

We use Phan's LDA Java implementation [*5] to run Latent Dirichlet Allocation on the collection. Dirichlet parameters α and β are set to $50/K$ and 0.1 respectively, where K is the number of latent topics. Previous researches [5], [6], [15] found that these values are suitable for broad categories of text data.

5.3.1 Evaluation Methodology

Ten people assessed the purity, novelty, and topic coverage of our proposed method. Because knowledge of the subject constitutes an important aspect in the quality of the evaluation itself, the evaluators were asked to read the collection composed of the Wikipedia article and the pages retrieved from the Web. After having knowledge related to the collection, they were asked to evaluate the sentences retrieved by our method. For purity and novelty, they were asked to determine if sentences were related thematically and also if they included novel information related to the Wikipedia article subject. For topic coverage, the evaluators were asked to select information from the collection of Web pages not included in the Wikipedia article but which they regarded as important, and to evaluate the information to ascertain if it was retrieved using our method.

5.3.2 Results

We present the averages of the values obtained for the purity, novelty, and topic coverage.

Purity

We obtained a good level of purity for the retrieved information; the average values for all the queries of more than 60% in most cases, as shown in **Fig. 2**. The retrieved sentences in most cases are identifiable by a human assessor as corresponding to a particular human-based topic with only a few sentences from other topics mixed. They are easily classified by users. Note that our purity metric provides a value of 1 when $s = 1$, because every selected latent topic is represented by this sentence only.

We conclude that although we used only a few pages for our evaluation, this fact did not constitute a major problem in retrieving well-grouped sentences.

Novelty

To estimate the effectiveness of our method for the retrieval of new information for each selected latent topic which we consider

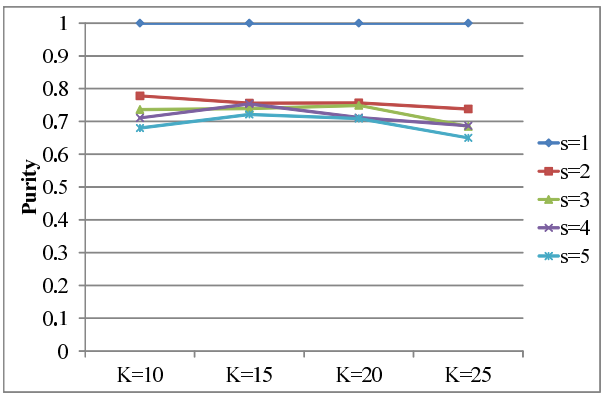

Fig. 2 Average purity.

[*5] http://jgibblda.sourceforge.net/

not well covered in the Wikipedia article, we compute the novelty over the data.

Figure 3 shows that, from $K = 15$ and above, more than 60% of the retrieved sentences are novel information. These results demonstrate that our method is effective in retrieving novel complementary information.

Topic Coverage

The coverage of all the topics in the collection constitutes one of our main goals. Good coverage of all the topics or aspects that are identifiable by humans as important to a subject frees users from a time-consuming process to obtain this complementary information.

In contrast to the purity and novelty, the topic coverage varies widely as the query changes. Therefore, we illustrate the topic coverage setting $s = 5$ for each of the ten queries varying K from 10 to 25 (**Fig. 4**). The IDs of the 10 queries are described in Table 1. For example, query 1 is "Yutaka Taniyama," query 2 is "Influvac." When $K = 10$, the topic coverage is relatively low, 0.58 on average. The number $|T|$ of selected latent topics becomes small if K is small, because $|T| \leq K$ holds. If $|T|$ is not large enough to cover the human-based topics for a query, then the topic coverage for the query becomes worse. Consequently, the topic coverage would be low if K is too small. For example, the worst result is obtained for query 6 and $K = 10$. On the other hand, our method setting $K \geq 15$ could find a large portion of the human-based topics not included in the Wikipedia articles. The average topic coverage is about 0.72 when $K = 15$, 0.78 when $K = 20$, and 0.89 when $K = 25$. Therefore, our method is effective for retrieving important information, from a human perspective, that is not included in the Wikipedia article.

Overall

From these results, we can conclude that our method is useful for users who seek complementary information from a collection of pages. For a user wondering if an article offers good coverage of the subject or desiring to have a quick general view, our method can be instrumental. Our method can also contribute greatly to helping Wikipedia contributors in upgrading Wikipedia articles because it retrieves important information that is not included in the article and presents it to users. It enables contributors to have a view of what is important in the collection without taking the time to read numerous pages, thereby saving time and effort. We found experimentally that a number of topics of $15 \leq K \leq 25$ provides good results.

5.4 Case Study

We present a case study for a query that is representative of a Wikipedia short and stub article.

5.4.1 Yutaka Taniyama

Purity

We present the results for the different values of K, the number of latent topics, and the number of retrieved sentences s. **Figure 5** shows that the sentences retrieved using our method are well grouped. The value of purity is higher than 60%, which is an acceptable range because the obtained results are easily classifiable by a human.

Novelty

Figure 6 illustrates the novelty of the sentences obtained by our method, varying K and s. For $K = 15$ and higher values,

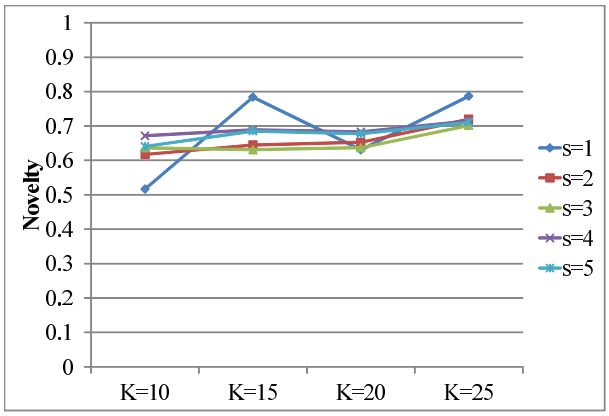

Fig. 3 Average novelty.

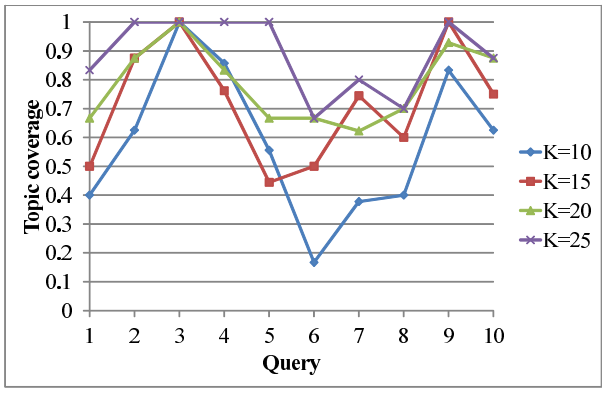

Fig. 4 Topic coverage for each query.

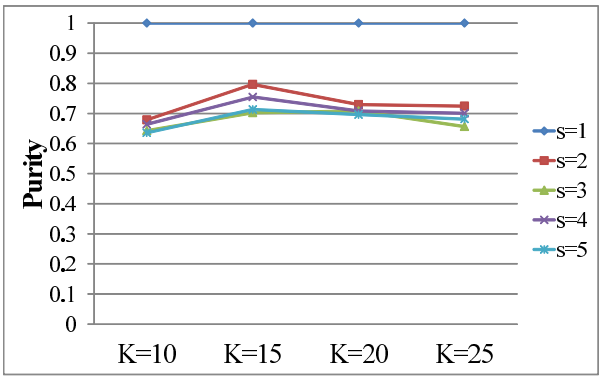

Fig. 5 Purity for Yutaka Taniyama.

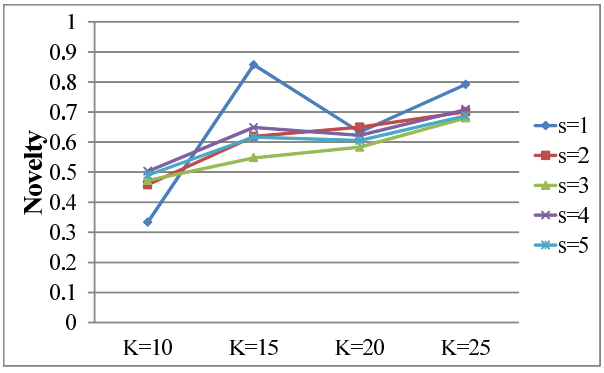

Fig. 6 Novelty for Yutaka Taniyama.

more than 50% of retrieved sentences were novel sentences. Our method is effective for retrieving new information.

Topic Coverage

First, we enumerate the human-based topics which evaluators regarded as important in the collection, but which are not fully covered in the Wikipedia article:

- Yutaka Taniyama's hobbies
- The fact that he was a sickly child.
- College life (University)
- Participation in a symposium
- Book
- Detailed information about his research (Effects, Contribution, Fermat Theorem)

The last human-based topic is partly covered in the Wikipedia article. Therefore, the information obtained from the Web for the topic is information improving topic coverage explained in Section 2. Because the others are not covered in the Wikipedia article, the obtained information for them is detailed information explained in Section 2.

We list below the retrieved sentences for two selected latent topics, named topics [a] and [b], obtained by our method setting $K = 20$ and $s = 5$.

Topic [a]:

- Yutaka Taniyama, whose insights ultimately engendered the solution, killed himself in 1958.
- Leonhard Euler, the greatest mathematician of the eighteenth century, had to admit defeat.
- Whole and colorful lives were devoted, and even sacrificed, in finding a proof.
- Sophie Germain took on the identity of a man to do research in a field forbidden to females, and made the most significant breakthrough of the nineteenth century.
- And then came Princeton professor Andrew Wiles, who had dreamed of proving Fermat's last theorem ever since he first read of it as a boy of ten in his local library.

Topic [b]:

- Taniyama's fame is mainly caused by two problems posed by him at the symposium on algebraic number theory held in Tokyo in 1955 (His meeting with Weil at this symposium was to have a major influence on Taniyama's work).
- Shimura later worked with Taniyama on this idea that modular forms and elliptic curves are linked and this form the basis of the Taniyama–Shimura conjecture: Every elliptic curve defined over the rational field is a factor of the Jacobian of a modular function field.
- Taniyama studied mathematics at the University of Tokyo after the end of World War II, and here he developed a friendship with another student named Goro Shimura.
- I don't quite understand it myself, but it is not the result of a particular incident, nor of a specific matter.
- Regarding the reason for taking his life he says: Until yesterday I had no definite intention of killing myself.

From Topic [a] we have information about other mathematicians who devoted their lives to solving a problem to which Taniyama's contribution constitutes a stepping stone to the solution. Furthermore, information is given about Andrew Wiles, who used Taniyama's result, was able to prove Fermat's last theorem. From Topic [b] we have information about the fact that Taniyama studied at the University of Tokyo and developed a friendship there with Shimura. We also have detailed information about his theory.

5.4.2 Influvac

In the case of "Influvac," the purity and the novelty are fairly high; the former is higher than 0.6 and the latter is higher than 0.8 when $K \geq 15$ and $s = 5$.

For topic coverage, the following human-based topics are covered in the collection:

- Description
- Composition
- Pharmacology, Action
- Dosage
- Administration
- Storage
- Effects (including general effects, interaction, before receiving Influvac, after receiving Influvac, children, elderly, pregnancy and lactation as human-based subtopics)

The Wikipedia article for "Influvac" explain "Description" topic only. Therefore, the information obtained from the Web for the topic is information improving topic coverage; the obtained information for the other topics is detailed information. In contrast to the Wikipedia article, our method was able to cover more than 90% of the human-based topics in the collection that are not contained in the Wikipedia article when $K = 20$ and $s = 5$. For example, our method found the following sentence for a latent topic corresponding to human-based subtopic "Effects for pregnancy": "For pregnant women with medical conditions that increase their risk of complications from the flu, administration of the vaccine is recommended, irrespective of their stage of pregnancy."

Similarly, our method found a number of sentences covering human-based topics not contained in Wikipedia for the other eight queries. Therefore, our method could complement information of Wikipedia effectively utilizing the Web.

5.5 Comparison with a Naïve Method

In order to confirm the effectiveness of our method, we compare the results of our method to those of a naïve method using tf-idf and the cosine similarity.

Given a query, the naïve method analyzes the result pages of our page-discriminative approach, as well as our method, explained in Section 4.1. Specifically, the naïve method first converts the result pages and the Wikipedia article into vectors using tf-idf. Then, this method computes the cosine similarity between the vector of the Wikipedia article and that of each result page, and chooses the top t pages dissimilar from the Wikipedia article as candidate pages containing information not included in the article. We set $t = 7$ throughout our experiment. The method then converts all sentences in each of the t pages into vectors using tf-idf. Finally, for each of the t pages, the method computes the cosine similarity between the vector of the page with that of each sentence in the page, and outputs the top s sentences similar to the page as sentences representative of the page, where s is the same parameter as that used in our method.

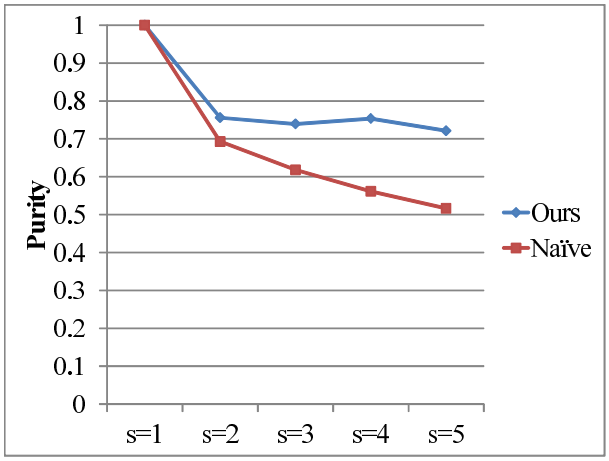

Fig. 7 Comparison of average purity.

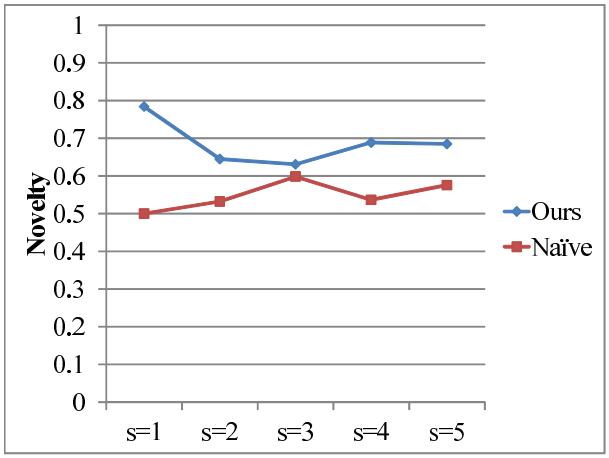

Fig. 8 Comparison of average novelty.

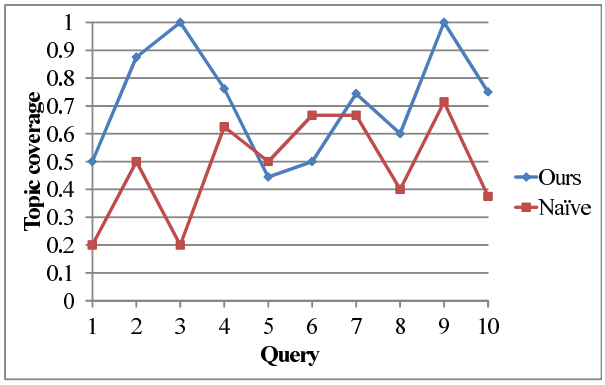

Fig. 9 Comparison of topic coverage.

We compare the results of our method setting $K = 15$ and those of the naïve method. **Figures 7** and **8** depict the results for purity and novelty, respectively. Recall that the purity must be 1 when $s = 1$. The purity of our method becomes superior to that of the naïve method as s increases. This fact indicates that the naïve method using tf-idf could not choose sentences of a common topic when s is large because the method does not consider topics. The novelty of our method is higher than that of the naïve method regardless the value of s.

Figure 9 illustrates the topic coverage of our method and the naïve method setting $s = 5$ for each of the ten queries. The indexes of the queries are the same as those used in Fig. 4. Our method achieve higher topic coverage than the naïve method does except two queries 5 and 6. Because our method setting $K > 15$ produces better results for these queries as depicted in Fig. 4, the small value $K = 15$ would be a reason of the low topic coverage as explained in Section 5.3.2. One of the interesting challenges in future work is to choose the optimal number of latent topics automatically. The average topic coverage of our method is 0.72, which is significantly higher than that of the naïve method, 0.48. The topic coverage of the naïve method could be fairly small, say less than 0.4. In contrast, the coverage of ours is higher than 0.4 for every query.

As a result, we confirm that our method outperforms the naïve method. Therefore, the proposed idea used in our method is effective in finding complementary information for Wikipedia from the Web.

6. Discussion

Retrieving complementary information, that is, new information or detailed information, is a difficult task because similar information can be translated differently in different articles according to the writer's language level, understanding of the subject, and so on. For example, in some articles, a piece of information can be summarized in one sentence, whereas in other articles the same information is extended to an entire paragraph. Another problem is the fact that information related to one topic can be dispersed throughout the article. The challenge of not flooding the user with unnecessary information is also important.

In our experiments, we mostly obtained the same range of value for the purity of the topics retrieved by LDA. However on a novelty level, we can observe a great variation of value according to the fact that the Wikipedia article included less information or not. For a Wikipedia article about a subject such as "Influvac" or "Plumbosolvency," which includes a small amount of information, mostly all retrieved information can be regarded as novel information.

Although we were able to retrieve novel information, duplicate information constitutes a problem because different selected topics can retrieve the same sentences. Aggregating the retrieved results on their similarity can be helpful in that matter.

Two major elements play an important role in the quality of the retrieved information. First, we have the number of pages we use as a dataset, also linked to the precedent element is the number of topics for the LDA process. Because different subjects can cover numerous topics and some just a few, choosing the number of topics can become an issue. The experiment results show improvement of the value of purity with an increase of the number of topics, but at $K = 15$ and above, little difference is apparent. We plan to conduct more experiments to elucidate this matter.

Although complementing Wikipedia articles from the Web presents many advantages, one important issue to consider is the fact that information related to the Web can be misleading in many aspects. Many studies [2], [3], [9] have been done to evaluate the trustworthiness of Wikipedia articles. Although we do not address this problem in this work, we consider the question interesting and plan to investigate it in future work.

7. Future Work

Our method for selecting the data considers specific words extracted from the Wikipedia article first sentence. It enables an efficient page discriminative process. However, the use of synonyms can help reduce the number of non-selected pages related to the topic. Query expansion techniques can also be used to widen the coverage of the related topic.

In the current state of our work, most of the parameters, especially the number of topics for the LDA process, are chosen manually. To have a fully automatic process, we plan to investigate Hierarchical Dirichlet Processes [16], which automatically infers the number of topics for a given collection of data.

The presentation of the results to the user constitutes an important task because we do not want the user to be overwhelmed with too much information. We plan to analyze ways to present information in a user friendly manner that enables the user to navigate easily through the complementary information. Finding a threshold for which information to present and which can not also be a way to regulate the amount of information that are presented to the user.

During the course of our evaluation, we had sentences including the same information as that in the Wikipedia article, but written in a different way, which were also retrieved. Removing these sentences from the extracted sentences helps reduce noise in the retrieved results. We plan to investigate how the structure of a Web page and the analysis of the textual content of the page coupled with natural language processing techniques can help us have an efficient complementary information extraction process.

Many techniques are useful to improve the quality of the topic retrieved by LDA such as removal of words that are not sense-bearing.

Our method is also useful for automatic link suggestions in Wikipedia because it enables the identification of topics and pages including novel information. The links that our method can provide are based not only on the relatedness of the article to the Wikipedia article but also on the knowledge contribution of the suggested links.

Wikipedia articles sometimes present trustworthiness and bias issues, especially for sensitive topics such as politics and history. In that matter, we plan to investigate how information retrieved from the Web can help in analyzing Wikipedia article trustworthiness and also bias.

8. Conclusion

We presented a method to complement the information included in a Wikipedia article with information retrieved from the Web. We defined complementary information of two types: information not included in the Wikipedia article and information providing more details related to a certain topic covered in the Wikipedia article. Our method, based on a probabilistic approach, takes into account the fact that a document can include multiple topics. Our experiments showed that our method was able to retrieve valuable complementary information. Therefore our method proved useful not only for the updating process of Wikipedia, but also to find pages that have more detailed information about a certain topic.

References

[1] Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet Allocation, *Proc. 15th Annual Conference on Neural Information Processing Systems* (*NIPS 2001*), pp.601–608 (2001).

[2] Chin, S.-C., Street, W.N., Srinivasan, P. and Eichmann, D.: Detecting Wikipedia Vandalism with Active Learning and Statistical Language Models, *Proc. 4th International Workshop on Information Credibility on the Web* (*WICOW*), pp.3–10 (2010).

[3] De la Calzada, G. and Dekhtyar, A.: On Measuring the Quality of Wikipedia Articles, *Proc. 4th International Workshop on Information Credibility on the Web* (*WICOW*), pp.11–18 (2010).

[4] Dean, J. and Henzinger, M.R.: Finding Related Web Pages in the World Wide Web, *Computer Networks*, Vol.31, No.11-16, pp.1467–1479 (1999).

[5] Denecke, K. and Brosowski, M.: Topic detection in noisy data sources, *The 5th International Conference on Digital Information Management* (*ICDIM*), pp.50–55 (2010).

[6] Griffiths, T.L. and Steyvers, M.: Finding scientific topics, *Proc. National Academy of Sciences*, 101 (Suppl. 1), pp.5228–5235 (2004).

[7] Eklou, D., Asano, Y. and Yoshiwaka, M.: How the Web can help Wikipedia: A Study on Information Complementation of Wikipedia by the Web, *Proc. 6th International Conference on Ubiquitous Information Management and Communication* (*ICUIMC*), No.9 (2012).

[8] Liu, B., Ma, Y. and Yu, P.S.: Discovering unexpected information from your competitors' web sites, *Proc. 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD*), pp.144–153 (2001).

[9] Lucassen, T. and Schraagen, J.M.: Trust in Wikipedia: How Users Trust Information from an Unknown Source, *Proc. 4th International Workshop on Information Credibility on the Web* (*WICOW*), pp.19–26 (2010).

[10] Ma, Q. and Tanaka, K.: Topic-Structure based Complementary Information Retrieval and Its Application, *ACM Trans. Asia Language Information Processing*, Vol.4, No.4, pp.475–503 (2005).

[11] Nadamoto, A., Aramaki, E., Abekawa, T. and Murakami, Y.: Content hole search in community-type content using Wikipedia, *Proc. 11th International Conference on Information Integration and Web-based Applications and Services* (*iiWAS*), pp.25–32 (2009).

[12] Nadamoto, A., Ma, Q. and Tanaka, K.: B-CWB: Bilingual Comparative Web Browser Based on Content-Synchronization and Viewpoint Retrieval, *World Wide Web Journal*, Vol.8, No.3, pp.347–367 (2005).

[13] Nadamoto, A. and Tanaka, K.: A comparative web browser (CWB) for browsing and comparing web pages, *Proc. 12th International World Wide Web Conference* (*WWW*), pp.727–735 (2003).

[14] Nakatani, M., Jatowt, A. and Tanaka, K.: Easiest-first search: Towards comprehension-based web search, *Proc. 18th ACM Conference on Information and Knowledge Management* (*CIKM*), pp.2057–2060 (2009).

[15] Steyvers, M. and Griffiths, T.: Probabilistic topic models, *Latent Semantic Analysis: A Road to Meaning*, Landauer, T., McNamara, S.D. and Kintsch, W. (Eds.), Laurence Erlbaum (2007).

[16] Teh, Y.W., Jordan, M.I., Beal, M.J. and Blei, D.M.: Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes, *Proc. 18th Annual Conference on Neural Information Processing Systems* (*NIPS 2004*), pp.1385–1392 (2005).

[17] Yeung, A., Duh, K. and Nagata, M.: Assisting Wikipedia Users in Cross-Lingual Editing, *WebDB Forum* (2010).

Damien Eklou received his B.E. in Computer Sciences from Gunma University in 2010 and his master degree in Informatics from Kyoto University in 2012. Currently, he is working at Deutsche Securities Inc. Japan as an Analyst in the Global Market Technology Division.

Yasuhito Asano received B.S., M.S. and D.S. in Information Science, the University of Tokyo in 1998, 2000 and 2003, respectively. In 2003–2005, he was a research associate of Graduate School of Information Sciences, Tohoku University. In 2006–2007, he was an assistant professor of Department of Information Sciences, Tokyo Denki University. He joined Kyoto University in 2008, and he is currently an associate professor of Graduate School of Informatics. His research interests include Web mining, network algorithms. He is a member of IEEE, DBSJ, and OR Soc. Japan.

Masatoshi Yoshikawa received his B.E., M.E. and Ph.D. degrees in Information Science from Kyoto University in 1980, 1982 and 1985, respectively. From 1985 to 1993, he was with Kyoto Sangyo University. In 1993, he joined Nara Institute of Science and Technology as an associate professor of Graduate School of Information Science. Currently, he is a Professor of Graduate School of Informatics, Kyoto University. His current research interests include XML information retrieval, databases on the Web, and multimedia databases. He is a member of ACM, IPSJ and IEICE.

(Editor in Charge: *Akira Maeda*)