Recommending Related Articles in Wikipedia via a Topic-Based Model

Wongkot Sriurai, Phayung Meesad, Choochart Haruechaiyasak

Department of Information Technology
Faculty of Information Technology
King Mongkut's University of Technology North Bangkok (KMUTNB)
1518 Pibulsongkarm Rd., Bangsue, Bangkok 10800

Department of Teacher Training in Electrical Engineering
Faculty of Technical Education
King Mongkut's University of Technology North Bangkok (KMUTNB)
1518 Pibulsongkarm Rd., Bangsue, Bangkok 10800

Human Language Technology Laboratory (HLT)
National Electronics and Computer Technology Center (NECTEC)
Thailand Science Park, Pathumthani 12120, Thailand

s4970290021@ kmutnb.ac.th
pym@kmutnb.ac.th
choochart.haruechaiyasak@nectec.or.th

Abstract: Wikipedia is currently the largest encyclopedia publicly available on the Web. In addition to keyword search and subject browsing, users may quickly access articles by following hyperlinks embedded within each article. The main drawback of this method is that some links to related articles could be missing from the current article. Also, a related article could not be inserted as a hyperlink if there is no term describing it within the current article. In this paper, we propose an approach for recommending related articles based on the Latent Dirichlet Allocation (LDA) algorithm. By applying the LDA on the anchor texts from each article, a set of diverse topics could be generated. An article can be represented as a probability distribution over this topic set. Two articles with similar topic distributions are considered conceptually related. We performed an experiment on the Wikipedia Selection for Schools which is a collection of 4,625 selected articles from the Wikipedia. Based on some initial evaluation, our proposed method could generate a set of recommended articles which are more relevant than the linked articles given on the test articles.

1 Introduction

Wikipedia is a well-known free-content encyclopedia written collaboratively by volunteers and sponsored by the non-profit Wikipedia Foundation[1].The aim of the project is to develop a free encyclopedia for many different languages. At present, there are over 2,400,000 articles available in English and many in other languages. The full volume of Wikipedia contents, however, contains some articles which are unsuitable for children. In May 2007, the SOS Children's Villages, the world's largest orphan charity, launched the Wikipedia Selection for Schools[2]. The collection contains 4,625 selected articles based on the UK National Curriculum and similar curricula elsewhere in the world. All articles in the collection have been cleaned up and checked for suitability for children.

The content of Wikipedia for Schools can be navigated by browsing on a pictorial subject index or a title word index of all topics. Table 1 lists the first-level subject categories available from the collection. Organizing articles into the subject category set provides users a convenient way to access the articles on the same subject. Each article contains many hypertext links to other articles which are related to the current article. However, the links which were assigned by the authors of the article cannot fully cover all related articles. One of the reasons is due to the fact that there is no term describing related articles within the current article.

Table 1: The subject categories under the Wikipedia Selection for Schools.

Category	**Articles**	**Category**	**Articles**
Art	74	Business Studies	88
Citizenship	224	Countries	220
Design and Technology	250	Everyday life	380
Geography	650	History	400
IT	64	Language and literature	196
Mathematics	45	Music	140
People	680	Religion	146
Science	1068		

[1]Wikipedia. http://en.wikipedia.org/wiki/WikiPedia

[2]Wikipedia Selection for Schools. http://schools-wikipedia.org

Some previous works have identified this problem as the missing link problem and also proposed some methods for automatically generating links to related articles. J. Voss [Vo05] presented an analysis of Wikipedia snapshot on March 2005. The study showed that Wikipedia links form a scale-free network and the distribution of in-degree and out-degree of Wikipedia pages follows a power law. S. Fissaha Adafre and M. de Rijke [FR05] presented an automated approach in finding related pages by exploring potential links in a wiki page. They proposed a method of discovering missing links in Wikipedia pages via a clustering approach.The clustering process is performed by grouping topically related pages using LTRank and then performing identification of link candidates by matching the anchor texts. Cosley et al. [Co07] presented SuggestBot, software that performs intelligent task routing (matching people with tasks) in Wikipedia. SuggestBot uses broadly applicable strategies of text analysis, collaborative filtering, and hyperlink following to recommend tasks.

In this paper, we propose a method for recommending related articles in Wikipedia based on the Latent Dirichlet Allocation (LDA) algorithm. We adopt the dot product computation for calculating the similarity between two topic distributions which represent two different articles. Using the proposed approach, we can find the relation between two articles and use this relation to recommend links for each article. The rest of paper is organized as follows. In Section 2, we describe the topic-based mode for article recommendation. Section 3 presents experiments and discussion. Finally, we conclude our work and put forward the directions of our future work in Section 4.

2 The Topic-Based Model for Article Recommendation

There have been many studies on discovering latent topics from text collections [SG06]. Latent Semantic Analysis (LSA) uses singular value decomposition (SVD) to map high-dimensional term-by-document matrix to a lower dimensional representation called latent semantic space [De90]. However, SVD is actually designed for normally-distributed data. Such a distribution is inappropriate for count data which is what a term-by-document matrix consists of. LSA has been applied to a wide variety of learning tasks, such as search and retrieval [De90] and classification [Bi08]. Although LSA have achieved important success but LSA have some drawbacks such as overfitting and inappropriate generative semantics [BNJ03].

Due to the drawbacks of the LSA, the Latent Dirichlet Allocation (LDA) has been introduced as a generative probabilistic model for a set of documents [BNJ03]. The basic idea behind this approach is that documents are represented as random mixtures over latent topics. Each topic is represented by a probability distribution over the terms. Each article is represented by a probability distribution over the topics. LDA has also been applied for identification of topics in a number of different areas. For example, LDA has been used to find scientific topics from abstracts of papers published in the proceedings of the national academy of sciences [GS04]. McCallum et al. [MC05] proposed an LDA-based approach to extract topics from social networks and applied it to a collection of 250,000 Enron emails. Newman et al. (2006) applied LDA to derive 400 topics such as Basketball, Harry Potter and Holidays from a corpus of 330,000 New York Times news articles and represent each news article as a mixture of these topics [Ne06].

Haruechaiyasak and Damrongrat [HD08] applied the LDA algorithm for recommending related articles in Wikipedia Selection for Schools, however, without providing any comparative evaluation.

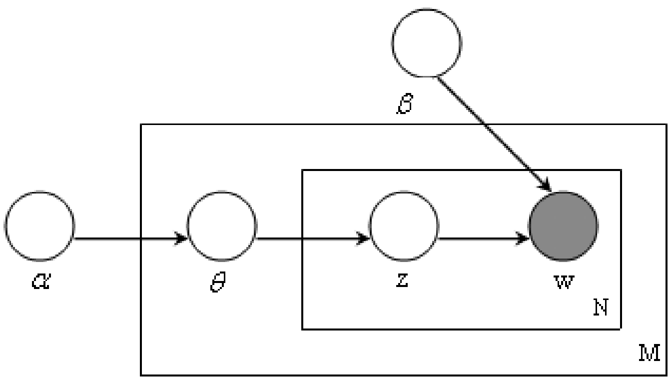

Figure 1: The Latent Dirichlet Allocation (LDA) model

Generally, an LDA model can be represented as a probabilistic graphical model as shown in Figure 2 [BNJ03]. There are three levels to the LDA representation. The variables α and β are the corpus-level parameters, which are assumed to be sampled during the process of generating a corpus. α is the parameter of the uniform Dirichlet prior on the per-document topic distributions. β is the parameter of the uniform Dirichlet prior on the per-topic word distribution. θ is a document-level variable, sampled once per document. Finally, the variables z and w are word-level variables and are sampled once for each word in each document. The variable N is the number of word tokens in a document and variable M is the number of documents.

The LDA model [BNJ03] introduces a set of K latent variables, called topics. Each word in the document is assumed to be generated by one of the topics. The generative process for each document w can be described as follows:

1. Choose $\theta \sim \mathrm{Dir}(\alpha)$: Choose a latent topics mixture vector θ from the Dirichlet distribution.
2. For each word $w_n \in w$
   (a) Choose a topic $z_n \sim \mathrm{Multinomial}(\theta)$: Choose a latent topic z_n from the multinomial distribution.
   (b) Choose a word w_n from $p(w_n | z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

In this paper, we focus on the Wikipedia Selection for schools for evaluating our proposed recommendation algorithm. Our proposed approach based on the topic model for recommending related articles and discovering missing links consists of three main processes as shown in Figure 2.

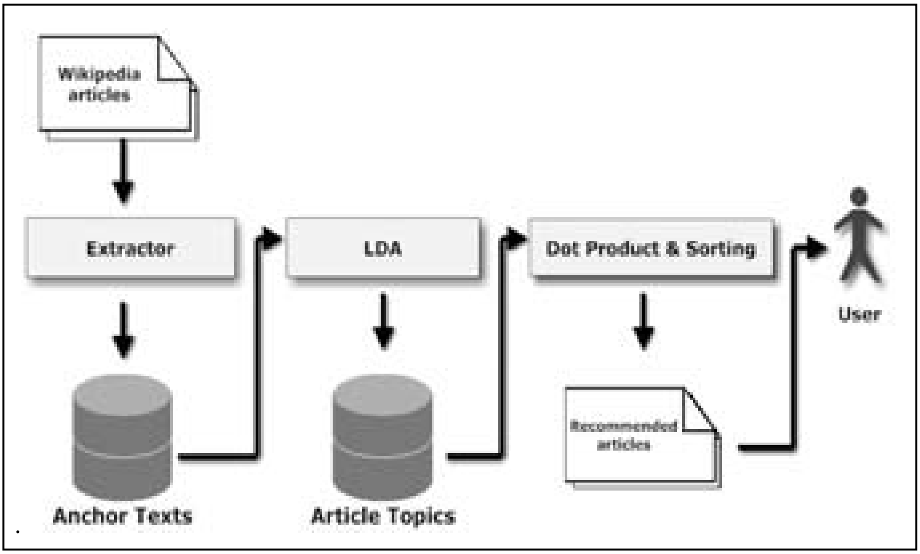

Figure 2: The proposed topic-based model via LDA algorithm for article recommendation.

1. Extract anchor-text links from all 4,625 Wikipedia Selection for School articles and store anchor texts in the database.
2. Prepare article titles and anchor texts from previous process as the input to generate the topic mode based on the LDA algorithm. The output from this step is the topic probability for each article.
3. The article similarity is computed by using the dot product between two topic probability vectors. The scores from the dot-product calculation are used to rank the top-10 articles that are related to the current article.

The process for recommending related articles can be explained in details as follows. The input data for the LDA algorithm consists of a document corpus. In this paper, we present each article with the title and anchor texts. The corpus is a set of m denoted by $D = \{d_0, ..., d_{m-1}\}$. Each document is a set of n topics denoted by $d_i = \{t_0, ..., t_{n-1}\}$. Finally, each topic is a set of distribution over p words denoted by $t_i = \{w_0, ..., w_{p-1}\}$.

To recommend related articles, we calculate the similarity between a given article and all other articles and select the ones with the highest similarity values. Given two articles represented as the topic distribution vectors, $d_i = \{t_0^i, ..., t_{n-1}^i\}$ and $d_j = \{t_0^j, ..., t_{n-1}^j\}$, the dot product can be calculated as follows.

$$d_i.d_j = \sum_{i=0}^{n-1} d_i d_j = t_1^i t_1^j + t_2^i t_2^j + ... + t_n^i t_n^j$$

3 Experiments and Discussion

The Wikipedia Selection for Schools is available from the SOS Children's Villages Web site[3]. We used the LDA algorithm provided by the linguistic analysis tool called LingPipe[4] to run our experiments. LingPipe is a suite of Java tools designed to perform linguistic analysis on natural language data. The tools are fast and robust enough to be used in a customer-facing commercial system. LingPipe's flexibility and included source make it appropriate for research use. LingPipe tools include a statistical named-entity detector, text classification and clustering. In this experiment, we apply the LDA algorithm provided under the LingPipe API and set the number of topics equal to 50 and the number of epochs to 2,000.

---

[3] SOS Children's Villages Web site. http://www.soschildrensvillages.org.uk/charity-news/wikipedia-for-schools.htm
[4] LingPipe. http://alias-i.com/lingpipe

Topic #14

Terms	Prob.
sun	0.041
planet	0.027
gravity	0.021
jupiter	0.017
mars	0.014
venus	0.014
hydrogen	0.014
mercury	0.013
solar	0.009
pluto	0.007

Topic #24

Terms	Prob.
oxygen	0.031
sodium	0.020
nitrogen	0.020
magnesium	0.016
potassium	0.014
copper	0.013
silicon	0.012
calcium	0.012
sulfur	0.011
mineral	0.011

Topic #27

Terms	Prob.
football	0.047
soccer	0.023
basketball	0.023
olympic	0.022
hockey	0.018
baseball	0.010
volleyball	0.008
sydney	0.006
tennis	0.004
sports	0.004

Topic #32

Terms	Prob.
art	0.041
paris	0.026
italy	0.026
rome	0.021
painting	0.019
leonardo	0.018
picasso	0.016
michelangelo	0.015
gallery	0.008
colour	0.006

Figure 3: Examples of topics generated by using the LDA algorithm.

Figure 3 shows some examples of topics generated by the LDA algorithm. Each table lists the top-10 terms ranked by the probabilistic values. It can be observed that the LDA could conceptually cluster highly similar terms into the same topics. For example, the terms art, gallery and painting are assigned into the same topic of 32. On the other hand, the topic 24 contains the terms related to the basic scientific elements and topic 27 contains the terms related to sports.

We applied the article recommendation approach described in the previous section on a sample set of articles. Figure 4 shows the comparison of the links within the article and the links from recommendation. The bold text shows recommended article links that not found in the article link made by human authors.

Article: Bill Clinton

Linked articles	Recommended articles
President of the U.S.	**Supreme Court of the U.S.**
United States Senate	President of the U.S.
John F. Kennedy	**John Tyler**
Vietnam War	**John Marshall**
House of Representatives	**William Henry Harrison**
Cuba	**Franklin D. Roosevelt**
Florida	House of Representatives
George W. Bush	**Benjamin Harrison**
Hurricane Katrina	**John W. Johnston**
Sydney	**American Civil War**

Article: Trigonometry

Linked articles	Recommended articles
Mathematics	Algebra
Sphere	**Algorithm**
Trigonometric functions	**Pi**
Science	Trigonometric functions
Calculus	**Computer science**
Programming language	**Arithmetic**
Ancient Egypt	**Topology**
Algebra	**Euclid**
Timur	**Prime number**
Electronics	**Applied mathematics**

Article: Mona Lisa

Linked articles	Recommended articles
16th century	**Drawing**
Oil painting	**History of painting**
Leonardo da Vinci	**Western painting**
Art	**Visual arts**
Government of France	**Painting**
Popular culture	**Michelangelo**
Andy Warhol	**Anthony van Dyck**
Elvis Presley	**Paul Cezanne**
Germany	Oil painting
Britney Spears	**Fine art**

Article: Cancer

Linked articles	Recommended articles
Cell	Cell
DNA	**Genetic code**
Proteins	**Life**
Viruses	**Genetics**
Latin	**DNA repair**
Hippocrates	**Punctuated equilibrium**
World War II	DNA
Brain	**Stroke**
Bird	**Sequence alignment**
Human	**Tay-Sachs disease**

Article: Dinosaur

Linked articles	Recommended articles
Vertebrates	Sauropodomorpha
Animals	**Oligocene**
Cretaceous	**Theropoda**
extinction	**Herrerasaurus**
pterosaurs	**Monoclonius**
Sauropodomorpha	Saurischia
Saurischia	**Camarasaurus**
Bird	Ornithischia
Ornithischia	**Protoceratops**
Heat	**Therizinosaurus**

Article: Television

Linked articles	Recommended articles
Telecommunication	**Electronics**
Broadcasting	**Central processing unit**
John Logie Baird	**Electrical engineering**
Latin	**Mass media**
Selenium	Telecommunication
Colour	**Communication**
BBC	**Publishing**
DVD	**Popular culture**
Video games	**Phonograph cylinder**
Sweden	**CPU cache**

Figure 4: Examples of article recommendation based on the topic-model approach.

The accuracy of the proposed recommendation approach is evaluated by the human assessor. The five assessors receive the article title, the linked articles and the recommended articles by our Topic-Based model. The assessor assigned the scores for each linked articles (LINK) and recommended articles (REC). The score is on the scale of 1 to 5. The average scores are shown in Table 2.

Table 2: Evaluation results between the linked articles (LINK) and the recommended articles (REC)

Article	**Score**	
	LINK	**REC**
Bill Clinton	2.275	3.675
Trigonometry	2.375	3.725
Mona Lisa	2.4	3.8
Television	2.125	3.125
Dinosaur	2.7	4.475
Cancer	2.075	3.475
Average	**2.325**	**3.7125**

The result shows that the scores from the recommended articles is higher than the scores from linked articles. This is especially true when the articles are about the definition of something and many articles are the class or specific type of that article, e.g., there are many dinosaur type articles that related to dinosaur definition article.

4 Conclusion and future works

Wikipedia is a well-known free-content encyclopedia. The content of Wikipedia can be navigated by browsing on a pictorial subject index or a title word index of all topics. Organizing articles into the subject category set provides users a convenient way to access the articles on the same subject. Each article contains many hypertext links to other articles which are related to the current article. However, the links which were assigned by the authors of the article cannot fully cover all related articles. One of the reasons is due to the fact that there is no term describing related articles within the current article. In this paper, we proposed a topic-model based method for recommending related articles in Wikipedia Selection for Schools. The topic model is generated by using the Latent Dirichlet Allocation (LDA) algorithm. The experimental results showed that the proposed method could help discover additional related articles, some of which are not listed as hyperlinks within a given article. The proposed recommend articles improve relevance score by 59.68%.

Our future works include the construction of an evaluation corpus. A set of random articles will be selected and all related articles will be judged by human experts. The corpus is useful in performing the empirical analysis of adjusting the LDA parameters. In this paper, we constructed the LDA model from textual information within the given articles. In our next work, we will extend the LDA model by including the neighboring information surrounding the current article. The neighboring information is, for example, the anchor texts of links into the current article. Using the neighboring information could provide richer and more coverage of information used to describe the current article.

References

[Bi08] Biro, I.; et al.: A Comparative Analysis of Latent Variable Models for Web Page Classification. Latin American Web Conference, pp. 23-28, 2008.

[BNJ03] Blei, D. M.; Ng, A. Y.; Jordan, M. I.: Latent dirichlet allocation. Journal of Machine Learning Research, 3(5): 993-1022, 2003.

[Co07] Cosley, D. et al.: SuggestBot: Using Intelligent Task Routing to Help People Find Work in Wikipedia: Proc. of the 12th acm international conference on intelligent user interfaces, New York, 2007.

[De90] Deerwester S. et al.: Indexing by latent semantic analysis. Journal of the American Society of Information Science, 41(6):391-407, 1990.

[FR05] Fissaha Adafre, S.; Rijke, M.: Discovering missing links in wikipedia: Proc. of the 3rd int. workshop on link discovery, 2005, pp. 90-97.

[GS04] Griffiths, T.; Steyvers, M.: Finding scientific topics: Proc. of the National Academy of Sciences, 2004, pp. 5228-5235.

[HD08] Haruechaiyasak, C; Damrongrat C.: Article Recommendation Based on a Topic Model for Wikipedia Selection for Schools: Proc. of the 11th International Conference on Asian Digital Libraries, 2008, S.339-342.

[MC05] McCallum, A.; Corrada-Emmanuel, A.; Wang, X.: Topic and role discovery in social networks: Proc. of IJCAI, 2005, pp. 786-791.

[Ne06] Newman, D. et al.: Analyzing entities and topics in news articles using statistical topic models. In Lecture Notes on Computer Science, Springer-Verlag, 2006.

[SG06] Steyvers, M.; Griffiths, T.: Probabilistic topic models, In T., Landauer, D., McNamara, S., Dennis, and W., Kintsch, (eds), Latent Semantic Analysis: A Road to Meaning, Laurence Erlbaum, 2006.

[Vo05] Voss, J.: Measuring Wikipedia: Proc. of Int. Conf. of the International Society for Scientometrics and Informetics, Stockholm, 2005.