Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media

Discovering Serendipitous Information from Wikipedia by Using its Network Structure

Yohei Noda†, Yoji Kiyota††, Hiroshi Nakagawa††

†Graduate School of Interdisciplinary Information Studies, University of Tokyo, Tokyo JAPAN
††Information Technology Center, University of Tokyo, Tokyo JAPAN
{noda,kiyota}@r.dl.itc.u tokyo.ac.jp
nakagawa@dl.itc.u tokyo.ac.jp

Abstract

Many researchers conducted studies on extracting relevant information from web documents. However, there are few studies on extracting serendipitous information. We propose methods to discover unexpected information from Wikipedia by using its network structure, for example, the distance between two categories. We evaluated two methods: a classification based method using support vector machines (SVMs), and a ranking based method using regression. We demonstrate advantages of regression over classification.

Introduction

Nowadays, a large number of people posts articles to weblog, microblogs (for example Twitter[1]), BBS, Q&A sites and wiki (including Wikipedia[2]). The number of consumer generated media (CGM) articles is sharply increasing. As these kinds of "collective intelligence" are accumulating in the web sphere, the demand of techniques to extract valuable and useful information from these articles is increasing.

Many studies have been conducted to extract information from the web sphere. These studies are mainly either "query-based" or "directory-based". Search engines help users to acquire information that they know at least its existence, and the resulting information is in line with what they expect. However, valuable information is not limited to information which users expect to be extracted. It also includes unexpected information which users do not know about it nor expect. The talent to detect unexpected or valuable things is called "serendipity." We apply the word "serendipitous information" to information that we cannot acquire without serendipity. We propose methods to help acquiring serendipitous information.

In this study, we extract serendipitous information from Wikipedia by using its network structure. The reasons why we use Wikipedia in this study are its scale, its characteristics of collective intelligence and its policies to maintain trustworthiness. Even though Wikipedia is not perfect in terms of trustworthiness, its benefits can easily be shown by the number of people who visits the site. Articles of Wikipedia are classified by its category system. By leveraging this character of Wikipedia, we can obtain serendipitous relations from its network.

The paper is structured as follows. We first introduce related work. Next, we describe our main idea on the overall relations between serendipitous information and the network of Wikipedia categories. We propose the method by using support vector machines (SVMs) and evaluate it. Then, we propose the method by using regression and evaluate it. Final section is conclusion and future work.

Related Work

In general, studies on recommendation system are evaluated from the viewpoint of precision of recommended items to the user's preferences. In addition to this evaluation figure, there are indication that unexpectedness and novelty should be included [1][2]. Recommendation systems, which adapt unexpectedness and novelty, have been shown to have high degree of satisfaction [3][4].

A study about content hole search by Nadamoto et al. [5] has a view of discovering unexpected knowledge. They proposed the method to get information which cannot be seen in the community-type contents for example, bulletin board systems (BBS) and social networking servics (SNS). They call unawareness information "the content hole", and proposed a method for discovering content hole by using Wikipedia.

Torishiki-kai is a study to look for some kinds of information ("trouble" or "how to") by using hyponymy relations which acquired from Wikipedia [6]. For example, the system extracts troubles and from web documents by using hyponymy relations of "trouble" and its synonyms.

Our Idea

Methods to get Information on the Web

According to Elain G.Toms, there are the following three methods to acquire information [7].

(1) from a search for information about a well-defined and known objects

[1] http://twitter.com/

[2] http://ja.wikipedia.org/

(2) *from a search for information about an object that cannot be fully described, but will be recognized on sight*

(3) *from accidental, incidental, or serendipitous*

In the web sphere, the method (1) and (2) are called "search", while the method (3) is called "web surfing". Serendipitous information is often acquired by (3), rather than (1) and (2). We cannot acquire serendipitous information by means other than (3). It is useful if serendipitous information were discovered from the web sphere. Our goal is to discover serendipitous information from Wikipedia systematically.

Structure and Characteristics of Wikipedia

We explain the structure and characteristics of Wikipedia. Each article in Wikipedia belong at least one category. Wikipedia Japanese edition has 9 main categories[3], and other categories are organized under the main categories. Categories and articles are linked to each other. Put simply, Wikipedia forms a graph structure consists of category nodes and article nodes.

We use relations between a Wikipedia article and two categories, as the basic unit of serendipitous information in Wikipedia In this paper, we call the basic unit as a "triple". We list up that kind of 1,637,472 triples from Wikipedia as candidacies and discovering serendipitous information from it. Figure.1 and Table.1 show examples of the triples.

Figure.1. An Example of a Triple

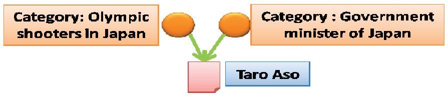

Category: Olympic shooters In Japan
Category : Government minister of Japan
Taro Aso

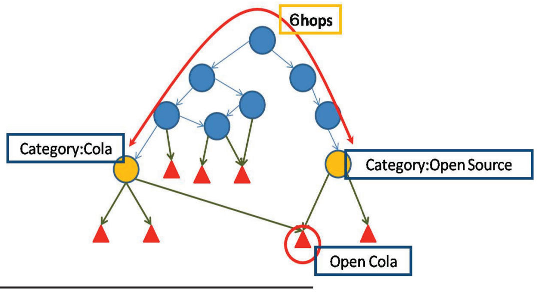
Table.1. Examples of the Candidates

Category A	Category B	Article
Government ministers of Japan	Olympic shooters of Japan	Taro Aso
Japanese actors	People from Tokyo	Takuya Kimura

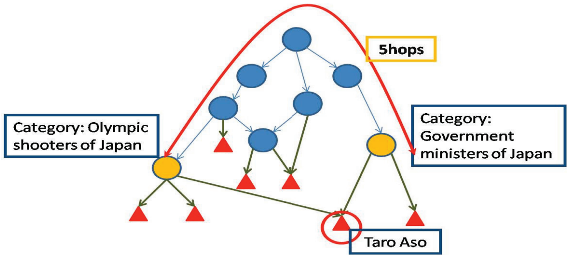
Serendipitous Information on the Wikipedia

Examples of Serendipitous Triples.

In the below example, the article "Open cola" belongs to both categories "Open Source" and "Cola". Natural flavorings in Coca-cola is a trade secret, however there is a project to make open sources soft-drink like Coca-cola.

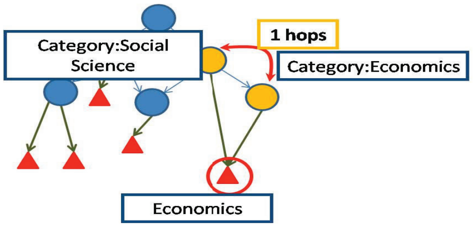


In the next example, the article "Taro Aso" belongs to both categories "Government minister of Japan" and "Olympic shooters of Japan". In general, every Japanese knows that "Taro Aso" was a government minister of Japan, however it is hard to find a person who experienced both the minister of Japan and an Olympic shooter. These two categories have only one common article, Aso Taro. This triple has rarity.

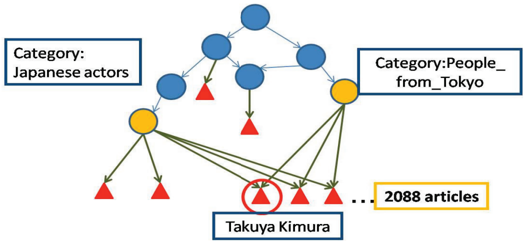

5hops
Category: Olympic shooters of Japan
Category: Government ministers of Japan
Taro Aso

Examples of Non-serendipitous Triples.

In the below example, the article "Economics" belongs to both categories "Social Science" and "Economics". However "Social Science" is a subcategory of "Economics". This is an obvious relation. So this triple is not serendipitous.



In the next example, 'Takuya Kimura" belongs to categories "Japanese actors" and "People from Tokyo". There are 2088 articles which belong to these two categories commonly. So this triple is a common practice. If the number of article that combines the set of categories were large, it is not serendipitous.



The above examples suggest a hypothesis that unexpectedness can be measured using the Wikipedia category networks.

Features

In this paper, we adopt the following four features, which express the Wikipedia category structure.

The Number of Child Articles in Category A or B. This feature describes how each category is general.

Hierarchical Level of Category A. This feature is the number of hops between the category A and the "root" category. The "root" category of Wikipedia Japanese

[3] Fundamental, Academia, Technology, Nature, Society, Geography, Humans, Culture and History

edition is called "Category: 主要カテゴリ (Category: Main categories)". The number of hops, in other words, the shortest path from the root category, describes how the category A is general.

The Number of common Article in Category A and B. This feature describes the rarity of the triple.

The Distance between two Categories. This feature describes the distance of the meaning. We count the hop of shortest path by using the Dijkstra algorithm [8]. We use an opensource software library "Jung" [9] to count this feature.

Supervised Data

We use the dump file of Wikipedia Japanese edition [10]. As a part of preprocessing step, we extracted the triples (Category A - Category B An article) by using an opensource software, Wik-IE [11].

Annotation was done by one person. He annotated the positive and negative labels to the triple if the article is unexpected from the view point of the relation between two categories. Table4 shows the distribution of the triples which are annotated both positive and negative.

Table4. The Distribution of the Labeled Triples

Feature	P/N	Mean	Standard Diviation
The number of child article in Category A	Posi	208.5	1126.3
	Nega	660.2	1952.1
Hierarchical level of Category A	Posi	3.9	5.6
	Nega	1.2	1.5
The number of child articles in Category B	Posi	119.2	649.3
	Nega	262.2	1682.5
Hierarchical level of Category B	Posi	4.3	5.7
	Nega	1.2	1.4
The number of common article of Category A & B	Posi	1.1	159.4
	Nega	0.8	373.9
Distance of Category A & B	Posi	4.8	3.8
	Nega	1.3	1.7

Discovering by Classification -SVM-

In this section, we propose the method to discover serendipitous information by using support vector machines (SVMs) [12].

Evaluation Setting

The evaluation data is based on the supervised data which we obtained at the previous section. We did 4 fold cross validation to the data. We tried linear kernel, polynomial kernel (quadratic and cubic) and Gaussian kernel. We use R [13] and its svm library kernlab [14].

Evaluation Result of SVMs

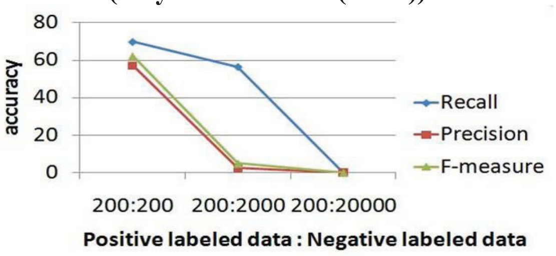
We describe the result of SVMs on Table6. Results have best accuracy in the cases of quadratic and cubic kernel. However, as is shown by the Figure.2, the more negative labeled data, the less accuracy. This is because the number of positive labeled data is quite a little. Because of the less of positive labeled data, it is difficult to classify by discriminative model, for example SVMs. So in the following section, we try to rank serendipitous triples by regression and evaluate by average precision.

Table6. Results by SVMs

Kernel	Recall	Precision	F-measure
Linear kernel	72.88%	54.43%	62.31%
Quadratic kernel	69.84%	57.14%	**62.86%**
Cubic kernel	73.33%	55.00%	**62.86%**
Gaussian kernel	76.36%	51.85%	61.76%

(positive:negative= 200:200)

Figure.2. Accuracy of Evaluation by SVM (Polynomial kernel (cubic))



Discovering by Regression

In this section, we propose the method to rank serendipitous triples and discover it from the ranked data by using regression. We use linear regression, nonlinear regression and logistic regression.

Evaluation Setting

We use supervised data as same as previous section. We use the labels of data, 1 or 0, as objective variables. For detail, we add a dimension to the supervised data for objective variable for regression.

We did nonlinear regression to the data and evaluated by 11-point average precision. We regard the objective variables as ranking barometer. It is hard to evaluate all 1,637,472 triples. So we fix one side of category by some categories and evaluated several cases. We chose two categories, "Category: マイケル・ジャクソン(Michael Jackson)" and "Category 地球温暖化(Global Warming)" triples for fixing one side of categories. The number of triples which was selected by fixing "Category:マイケル・ジャクソン (Michael Jackson)" is 44. While, the number of triples which was "selected by fixing "Category: 地球温暖化(Global Warming)" is 64.

Evaluation Result of Regression

Table7 shows the result of cubic regression. We could hardly say that average precision is high. However when the number of negative labeled data is ramped up, the accuracy is more stable than the case of SVM. Additionally, we suppose that relearning, by using user's annotation, make the accuracy better. This is because that the number of supervised data is still quite a little.

Table7. Average precision

positive : negative	200:200	200:2000	200:20000
Category:マイケル・ジャクソン(Michael Jackson)	0.190	0.239	0.236
Category: 地球温暖化 (Global Warming)	0.420	0.373	0.380

We got several serendipitous triples from Wikipedia category networks by using this method. Table. 8 and 9 show examples of serendipitous triples which appeared in evaluation data.

Table8. Examples of Serendipitous Triples under the "Category: Michael Jackson"(200:20000)

CategoryA	CategoryB	Article	Rank
LGBT の人物 (LGBT person)	マイケル・ジャクソン (Michael Jackson)	リサ・マリー・プレスリー (Lisa Marie Presley)	3
マイケル・ジャクソン (Michael Jackson)	著名な動物 (Famous animals)	バブルス (Bubbles)	10

Table9. Examples of Serendipitous Triples under the "Catgory: Global Warming"(200:20000)

CategoryA	CategoryB	Article	Rank
都市伝説 (Urban legend)	地球温暖化 (Global warming)	ワールド・ジャンプ・デー (World Jump Day)	1
地球温暖化 (Global warming)	ドキュメンタリー番組 (Documentaly television films)	地球温暖化詐欺 (映画) (The Great Global Warming Swindle)	9

Conclusion and Future Work

Conclusion

In this study, we propose the method to discovery serendipitous information from Wikipedia by using SVMs and regression. The result shows that the ranking-based method using regression is more appropriate to this problem. The accuracy is low, but we expect that the more positive labeled data acquired by using this system, the more accuracy increased.

Future Work

In the next research step, we plan to acquire a large number of labeled data deemed positive by user's feedback by using the methods. Relearning the model using abundant supervised data should provide us with further insights into discovering serendipitous information. Then we would like to do relearning by using those abundant supervised data.

References

[1] Murakami, T. and Mori, K. and Orihara, R.: Metrics for evaluating the serendipity of recommendation lists, Lecture Notes in Computer Science, 4914,pp.40 46, 2008

[2] Swearingen, K. and Sinha, R.: Beyond algorithms: An HCI perspective on recommender systems, ACM SIGIR 2001Workshop on Recommender Systems,2001

[3] McNee, S.M. and Riedl, J. and Konstan, J.A.:Making recommendations better: an analytic model for human-recommender interaction, CHI'06 extended abstracts on Human factors in computing systems, pp.1108,2006

[4] Murakami, T. and Mori, K. and Orihara, R.:A Method to Enhance Serendipity in Recommendation and its Evaluation,Transactions of the Japanese Society for Artificial Intelligence, vol.24,pp.428--436,2009

[5] Nadamoto, A. and Aramaki, E. and Abekawa, T. and Murakami, Y.:Content hole search in community-type content, Proceedings of the 18th international conference on World wide web, pp.1223--1224,2009

[6] Shinzato, K. and Torisawa, K.:Acquiring hyponymy relations from web documents, Proceedings of HLT-NAACL, vol.80, 2004

[7] Toms, E.G.: Serendipitous information retrieval, Proceedings of the First DELOS Network of Excellence Workshop on Information Seeking,pp.11 12, 2000

[8] Dijkstra, EW.:A note on two problems in connexion with graphs, Numerische mathematik, vol.1(1), pp.269--271, 1959

[9] JUNG, software library available at http://jung.sourceforge.net/

[10] Wikipedia dump data, data available at http://download.wikimedia.org/

[11] Wikipedia data analysis tool, Wik-IE, sourceforge, software available at http://wik-ie.sourceforge.jp/

[12] Vapnik, V. Statistical learning theory, Wiley, New York, 1998

[13] R, Application available at http://www.r-project.org/

[14] CRAN Package kenrlab, Package available at http://cran.r-project.org/web/packages/kernlab/index.html