Discovering Unexpected Information on the basis of Popularity/Unpopularity Analysis of Coordinate Objects and their Relationships

Kosetsu Tsukuda
Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo, Kyoto, Japan
tsukuda@dl.kuis.kyoto-u.ac.jp

Hiroaki Ohshima
Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo, Kyoto, Japan
ohshima@dl.kuis.kyoto-u.ac.jp

Mitsuo Yamamoto
Denso IT Laboratory
Shibuya-ku, Tokyo, Japan
miyamamoto@d-itlab.co.jp

Hirotoshi Iwasaki
Denso IT Laboratory
Shibuya-ku, Tokyo, Japan
hiwasaki@d-itlab.co.jp

Katsumi Tanaka
Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo, Kyoto, Japan
tanaka@dl.kuis.kyoto-u.ac.jp

ABSTRACT

Although many studies have addressed the problem of finding Web pages seeking relevant and popular information from a query, very few have focused on the discovery of unexpected information. This paper provides and evaluates methods for discovering unexpected information for a keyword query. For example, if the user inputs "Michael Jackson," our system first discovers the unexpected related term "karate" and then returns the unexpected information "Michael Jackson is good at karate." Discovering unexpected information is useful in many situations. For example, when a user is browsing a news article on the Web, unexpected information about a person associated with the article can pique the user's interest. If a user is sightseeing or driving, providing unexpected, additional information about a building or the region is also useful. Our approach collects terms related to a keyword query and evaluates the degree of unexpectedness of each related term for the query on the basis of (i) the relationships of coordinate terms of both the keyword query and related terms, and (ii) the degree of popularity of each related term. Experimental results show that considering these two factors are effective for discovering unexpected information.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval—*Information Search and Retrieval*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SAC'13 March 18-22, 2013, Coimbra, Portugal.
Copyright 2013 ACM 978-1-4503-1656-9/13/03 ...$10.00.

General Terms

Algorithms, Experimentation

Keywords

Coordinate term, Wikipedia, unexpected information

1. INTRODUCTION

Search engines such as Google[1], Yahoo[2], and Bing[3] return search results ranked by relevance and popularity relative to the input query. In most cases, higher ranked Web pages include more relevant and popular information. Some research has proposed innovative methods for documents retrieval. For example, BM25 [18] has been proposed as a state-of-the-art text-based ranking function, and HITS [11] and PageRank [3] are link-based ranking algorithms. On the basis of these studies, many additional studies have reported improved methods for the retrieval of appropriate query results [6, 8, 9, 20].

A disadvantage of these studies is that they do not address unexpected information. To the best of our knowledge, there have been very few studies that focus on discovering unexpected information on the Web [12, 14, 15], although there has been a great deal of research focused on extracting unexpected or unusual frequent rules in the field of data mining [2, 16, 17]. When a user queries a search engine, the retrieved Web pages contain a wide variety of information relative to the query. These pages can contain details ranging from well-known to unexpected information. For example, for the query "Michael Jackson," it is well known that "Michael Jackson is a singer," but it is generally unknown that "Michael Jackson is good at karate." The user can find common known information about a query easily because it is often included in the top ranked search engine result pages (SERP); however, comparatively less known information would likely appear in lower ranked Web pages. Even

[1] http://www.google.com

[2] http://www.yahoo.com

[3] http://www.bing.com

if top ranked Web pages include unexpected information, it is usually buried in a lot of other information and is difficult for the user to find.

Discovering relevant unexpected information relative to a keyword query is useful in certain situations. For instance, when a user searches the Web for information about a specific person, finding unexpected information can pique the user's interest. Similarly, if unexpected information about a person or incident is displayed when a user is browsing a news article, the information can also pique the users' interest. Moreover, when a user is sightseeing or driving, showing unexpected information about a building or the surrounding area may come in handy. Hence our objective is to discover unexpected information relative to keywords, such as specific people, facilities, or regions.

In this research, we target information that contains two objects. For example, in the information "Michael Jackson is a singer," one object is "Michael Jackson" and the other is "singer." We denote an object given as a keyword query as a "theme term" and an object that is related to the theme term as a "related term." Detailed explanations of theme terms and related terms are provided in Section 3. Our approach has three steps. First, given a query keyword (theme term) q, we collect its related terms $L_q = \{e_1, e_2, \cdots e_n\}$. We use Wikipedia[4] to collect a very large set of related terms. Next, we evaluate the degree of unexpectedness of each related term e_i for q on the basis of relationships of coordinate terms of q and e_i, and the degree of popularity of e_i. We hypothesize that when the objects are popular but the relationship between the objects is unpopular, the information is unexpected. In our method, we utilize the link structure between terms in Wikipedia and the super-sub relation between terms. Finally, we extract a sentence from a Wikipedia article that includes a related term with a high degree of unexpectedness and present it to a user as unexpected information.

We conducted an experiment using 75 queries in five domains: the names of people, regions, products, facilities, and organizations. Our results show the effectiveness of our algorithm considering the popularity of related terms of a theme term and the unpopularity of the term-relationships.

The remainder of this paper is organized as follows. Section 2 surveys related work. Section 3 explains the hypothesis of unexpected information as used in this research. Section 4 proposes methods for calculating the degree of unexpectedness for each related term for a query. Section 5 describes the experimental set-up and reports the results. A summary of the research and plans for future studies are presented in Section 6.

2. RELATED WORK

In the field of information retrieval, there are many methods for returning a list of search results ranked by relevance and popularity [3, 6, 8, 9, 11, 18, 20]. In most cases, the top ranked documents tend to include relevant and popular (well-known, major) information for the query. However, finding unexpected information by applying these methods is difficult.

In the field of information extraction, many studies have proposed techniques for finding useful knowledge from the Web [1, 4, 7]. Some studies used machine learning [4, 5], while others utilize syntactical pattern matching [7] or a bootstrapping technique [1]. These studies address discovering generally known and popular information rather than unexpected information because the objective is to create a computer-understandable knowledge base to realize much more effective retrieval of Web information.

To the best of our knowledge, very few studies focus on discovering unexpected information [12, 14, 15]. Noda et al.[15] used a relationship between categories in Wikipedia to discover unexpected knowledge. Using their method, a user can find that "Taro Aso" belongs to the category "Japan's premier" and to the category "participant in an Olympic shooting event." Only Taro Aso belongs to the two categories, and the fact "Taro Aso is a Japan's premier and a participant in an Olympic shooting event." is unexpected. In Wikipedia, articles do not belong to many categories, and therefore, their approach is limited. Nadamoto et al.[14] proposed a method for searching for a user's unawareness of information in community-type content, such as blogs and social networking services. They refer to such information as a "content hole" and define seven types of content holes [13]. Liu et al.[12] proposed methods to help a company find unexpected information from competitors' Web sites by comparing their Web sites with that of the competitors. This approach compares sites for information such as important keywords and outgoing links and displays the differences to the user. Their objective was to discover unexpected information that is not included in a particular Web site or bulletin board system. Our objective is to discover unexpected information for a keyword for an unrestricted search.

In the field of association rule mining, the frequency-based rule for discovering information becomes less interesting because most frequency rules are obvious. Instead, discovering unexpected patterns has received increasing attention [17, 16, 2]. However, we cannot use these approaches because they involve well-structured rules and have clear syntax and semantics. Because information on the Web is not fully structured, a different approach is required to discover unexpected information from the Web.

3. UNEXPECTED INFORMATION

We target information that contains two objects. Here an object is an essential element that, when combined with another essential element (object), constructs the information. For example, in the case of the information "Michael Jackson is good at karate," "Michael Jackson" and "karate" are objects because they are important elements. This information could be shown when the user conducts a Web search with the query "Michael Jackson," or browses a news article about "Michael Jackson." In these situations, we find unexpected information about the input keyword "Michael Jackson." We denote the object that is given as an input keyword as a "theme term," and we refer to an object related to a theme term as a "related term." There are various types of related terms for the theme term "Michael Jackson," for example "singer," "THIS IS IT," "Stevie Wonder," among many others.

When two terms have a common hypernym, they are coordinate terms. For instance, "Michael Jackson" and "Stevie Wonder" are coordinate terms because they have a common hypernym, "singer." "Michael Jackson" and "Maria Sharapova" are also coordinate terms because of the common hypernym, "human beings." However, "Stevie Wonder" is an appropriate coordinate term because "Michael Jackson" and "Stevie Wonder" have many common hypernyms in addition to "singer," for example "male" and "vegetarian." On the other hand, "Maria Sharapova" is a less-appropriate coordinate term. There are degrees of difference among the coordinate terms of a theme term. In this paper, we denote a term that has many hypernyms with a term as an "appropriate coordinate term," and a term that has few hypernyms with a term as a "less-appropriate coordinate term."

To describe the type of information people perceive unexpected as something relative to the theme term, its related term, and their coordinate terms, we inspect four examples, each with the theme term "Michael Jackson." The information "Michael Jackson won a

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

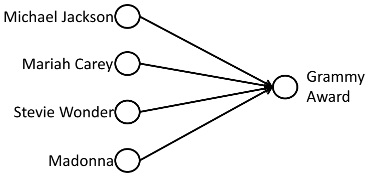

Figure 1: Related term "Grammy Award" is also related to appropriate coordinate terms of "Michael Jackson."

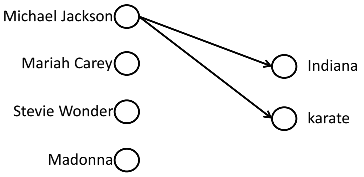

Figure 2: Related terms "Indiana" and "karate" are not related to appropriate coordinate terms of "Michael Jackson."

Grammy Award" is not unexpected to most people as many prominent artists have won the award. That is, appropriate coordinate terms of "Michael Jackson" also have the term "Grammy Award" as a related term (See Figure 1). For the information "Michael Jackson is from Indiana" and "Michael Jackson is good at karate," appropriate coordinate terms of "Michael Jackson" may not have "Indiana" or "karate" as related terms, as is shown in Figure 2. Although these two examples have the same structure, the information "Michael Jackson is from Indiana" may not be common knowledge but it is not entirely unexpected. All singers are from a certain place; therefore, this information is just an example of the same. That is, appropriate coordinate terms of "Michael Jackson" have appropriate coordinate terms of "Indiana" as their related terms (See Figure 3). In contrast, most people do not expect Euro-American singers to be good at Japanese martial arts. Consequently, the degree of unexpectedness of the information "Michael Jackson is good at karate," is quite high. That is, appropriate coordinate terms of "Michael Jackson" do not have appropriate coordinate terms of "karate" as their related terms (See Figure 4). We also consider the information "Michael Jackson bought an R-360." Here, "R-360" is a game machine. In this case, appropriate coordinate terms of "Michael Jackson" do not have appropriate coordinate terms of "R-360" as their related terms. Although this information has the same structure as "Michael Jackson is good at karate," the degree of unexpectedness of this information would be low because the term "R-360" is not generally known and therefore the degree of popularity is low. That is, we hypothesize that people do not perceive information as unexpected if it includes an unknown related term. Therefore, we also need to consider the degree of popularity of each related term.

In our approach, we regard unexpected information as information in which coordinate terms of a theme term do not have a relationship with a well-recognized related term and its coordinate terms. Given a theme term q and its related term e_i, we define a function $Rel(q, e_i)$ that represents the degree of relationship between q and e_i. The function $Cog(e_i)$ represents the popularity degree of e_i. We then define a function f that combines these functions and calculate the degree of unexpectedness of the pair of q and e_i: $Unexp(q, e_i) = f(Rel(q, e_i), Cog(e_i))$.

4. CALCULATING THE DEGREE OF UNEXPECTEDNESS

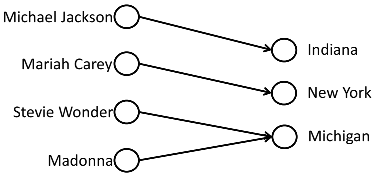

Figure 3: appropriate coordinate terms of "Michael Jackson" include appropriate coordinate terms of "Indiana" as a related term.

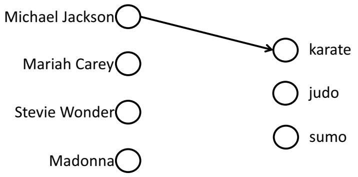

Figure 4: appropriate coordinate terms of "Michael Jackson" do not include appropriate coordinate terms of "karate" as a related term.

Given a theme term, the degree of unexpectedness of each related term is calculated as follows:

1. Collect a set of related terms $L_q = \{e_1, e_2, \cdots e_n\}$ for a theme term q.
2. Collect hypernyms and coordinate terms of q and those of each related term.
3. Calculate the strength of a relationship $Rel(q, e_i)$ between q and each related term.
4. Calculate the degree of popularity $Cog(e_i)$ for each related term.
5. Calculate the degree of unexpectedness $Unexp(q, e_i)$ of each related term for q.

In the following subsections, we explain each step in detail.

4.1 Collecting a set of related terms

In this paper, we regard anchor texts in a Wikipedia article of the theme term q as the related terms for q. Anchor texts are used to link related Wikipedia articles. In the case of "Michael Jackson," there are a total of 819 anchor texts; for example, "Thriller," "Paul McCartney" and "PlayStation 3," all appear as anchor texts. We focus on Wikipedia articles for three reasons. The first reason is that there are fewer noise terms in Wikipedia articles as they generally focus on information about a theme term q. The second reason is that Wikipedia articles primarily contain objective information. We are not targeting unexpected information derived from personal opinions or impressions; we are only interested in information written from an objective perspective. The third reason is that, as a matter of policy, Wikipedia does not link to a term if the term is not directly related to the title of an article. Therefore, we collect all Wikipedia anchor texts in an article of q as related terms for q.

4.2 Collecting hypernyms and coordinate terms

To collect coordinate terms, we used an open source "hypernym/hyponym extraction tool,"[5] on the Japanese Wikipedia. This

[5] http://nlpwww.nict.go.jp/hyponymy/index.html

tool contains $223,772$ hypernyms and $2,751,046$ hyponyms. These hierarchized terms are category names and nouns that occur in the titles of Wikipedia articles. Using this data, we can easily extract hypernyms of a term and coordinate terms that have hypernyms in common with the term. For instance, “Michael Jackson” has a total of 69 hypernyms such as “singer” and “Guinness world record holder.” If a term has at least one common hypernym with “Michael Jackson,” the term is a coordinate term of “Michael Jackson.”

4.3 Calculating the degree of relevance between a subject term and its related term

Before explaining our proposed method in detail, we will describe it visually. In Figure 5, vertices represent terms and the edges represent their relationships. The graph is constructed from the following vertices. We denote the set of hypernyms of term t with $hyper(t)$, the set of hyponyms of t with $hypo(t)$, and the set of related terms of t with $rel(t)$.

- $Q = \{q\}$ĄD
- $H_q = \{x | x \in hyper(q)\}$ĄD
- $C_q = \{x | x \in hypo(y), y \in H_q, x \notin Q\}$ĄD
- $L_q = \{x | x \in rel(q)\}$ĄD
- $H_{lq} = \{x | x \in hyper(y), y \in L_q\}$ĄD
- $L_c = \{x | x \in rel(y), y \in C_q, x \notin L_q\}$ĄD

In Figure 5, the black circle, white circle, black triangle, white triangle vertices represent a term in Q, C_q, L_q, and L_c, respectively. A square vertex represents a term in H_q or H_{lq}.

Edges exist between two terms if and only if one term is a hypernym of the other term or one term is a related term of the other term. In the following, (n_1, n_2) means that there is an edge between a vertex n_1 and a vertex n_2.

- (q, x) where $x \in H_q$ĄD
- (x, y) where $x \in H_q$, $y \in C_q$, and $y = hypo(x)$ĄD
- (x, y) where $x \in C_q$, $y \in L_c$, and $y = rel(x)$ĄD
- (x, y) where $x \in C_q$, $y \in L_q$, and $y = rel(x)$ĄD
- (x, y) where $x \in L_c$, $y \in H_{lq}$, and $y = hyper(x)$ĄD
- (x, y) where $x \in H_{lq}$, $y \in L_q$, and $x = hyper(y)$ĄD

This graph does not include edges between the theme term and its related terms because the objective is to demonstrate the ease in reaching all related terms of a theme term. That is, we assume that if it is *easy* to reach a specific related term from a theme term, the related term is expected. As we indicated previously, the term “Grammy Award” is not unexpected for “Michael Jackson.” As shown in Figure 5, there are many paths to reach “Grammy Award” from “Michael Jackson” through appropriate coordinate terms such as “Michael Jackson → singer → Stevie Wonder → Grammy Award” and “Michael Jackson → singer → Mariah Carey → Grammy Award.” In this case, we think it is *easy* to reach the related term. In the case of “Indiana,” there may be very few paths to directly reach “Indiana” in one step from appropriate coordinate terms of “Michael Jackson.” However, there are many paths to reach “Indiana” from appropriate coordinate terms through the hypernyms of “Indiana” such as “Michael Jackson → singer → Stevie Wonder → Michigan → American state → Indiana” and “Michael Jackson → member of USA for Africa → Diana Ross → Michigan → place name → Indiana.” On the other hand, there are no paths to reach “karate” from appropriate coordinate terms, even through the hypernyms of “karate.” Although there may be a few paths from less-appropriate coordinate terms directly or through the hypernyms, we assume that it is *difficult* to reach “karate” from “Michael Jackson” and that there is potential for an unexpected term; for example “Michael Jackson → human beings → Jean Cocteau → sumo → Japanese sport → karate.”

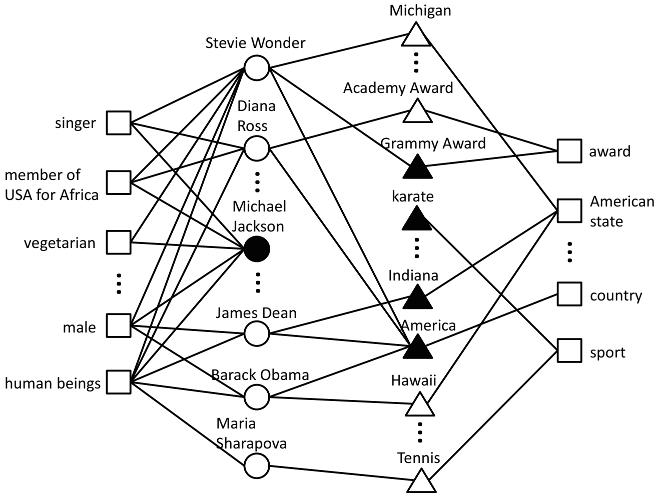

Figure 5: An example of the graph for a theme term “Michael Jackson.”

To evaluate the strength of the relationship between a theme term and each of its related terms, we first construct a graph as described above. We regard the presence or absence of the relationship between a theme term and its related term as the presence or absence of a path between these two terms. We calculate the degree of association of the pair by considering the strength of the relationship between them as the ease in reaching the related term from a theme term. The more difficult it is to reach a related term from a theme term, the lower the degree of association.

In the next subsections, we divide the graph into three subgraphs and evaluate the degree of association of related terms.

4.3.1 Calculating the degree of coordination to a theme term

First, we consider a bipartite graph $G_1 = (Q \cup C_q \cup H_q, E_1)$ that is constructed from q, its hypernyms, and their hyponyms. Here E_1 is a set of edges between H_q and $Q \cup C_q$. An edge exists between $h_i \in H_q$ and $t_j \in Q \cup C_q$ when h_i is a hypernym of t_j.

We apply the Co-HITS algorithm [6] to the bipartite graph and calculate the degree of coordination to q for each term in C_q. The Co-HITS algorithm is an expanded version of the HITS algorithm [11]. In the HITS algorithm, a Web page that provides important information is called an *authority*, and a Web page that links to important authorities is called a *hub*. A good hub is a page that points to many good authorities, and a good authority is a page that is pointed to by many good hubs. In our bipartite graph, a hypernym and a hyponym correspond to a hub and an authority, respectively. We denote the hub score of h_i and the authority score of t_j with x_i and y_j, respectively, and calculate these scores as follows:

$$x_i = \sum_{t_j \in Q \cup C_q} w_{ji}^{th} y_j, \quad y_j = \sum_{h_i \in H_q} w_{ij}^{ht} x_i, \qquad (1)$$

where w_{ji}^{th} and w_{ij}^{ht} represent the weight of edges, and w_{ji}^{th} represents the transition probability from t_j to h_i. If we apply the HITS algorithm to the bipartite graph G_1, vertices that have a very large number of hyponyms, for example "human beings," have a high score. Hence each hyponym of "human beings" has a misleading high score, and terms sharing hypernyms that have many hyponyms become appropriate coordinate terms of q. To solve this problem, we use the Co-HITS algorithm expressed as follows:

$$x_i = (1 - \lambda_h)x_i^0 + \lambda_h \sum_{t_j \in Q \cup C_q} w_{ji}^{th} y_j, \qquad (2)$$

$$y_j = (1 - \lambda_t)y_j^0 + \lambda_t \sum_{h_i \in H_q} w_{ij}^{ht} x_i, \qquad (3)$$

where x_i^0 and y_j^0 represent the initial scores for terms h_i and t_j, respectively. In the Co-HITS algorithm, the more edges a vertex has, the smaller the weights of the edges become. Specifically the weight of the edge from h_i to t_j is represented by $w_{ij}^{ht} = \frac{1}{|hypo(h_i)|}$, and the weight of the edge from t_j to h_i is represented by $w_{ji}^{th} = \frac{1}{|hyper(t_j)|}$. Moreover, the Co-HITS algorithm considers the initial score of each vertex. We set the initial value of q as 1 and the initial values of the remaining vertices as 0 because the objective of applying the Co-HITS algorithm is to calculate the degree of coordination to q. The parameters are $\lambda_h \in [0, 1]$ and $\lambda_t \in [0, 1]$. If we regard an initial score as important, we set their values close to 0. We set $\lambda_h = \lambda_t = 1$ because only the vertex of q has an initial score and the results did not change significantly when the values of the parameters were changed. We regard the convergent score of y_j as the degree of coordination of t_j to q.

4.3.2 *Calculating the degree of relevance between a theme term and each related term (1)*

To calculate the degree of relevance between a theme term and each of its related terms, we consider a graph G_2 that includes all vertices in C_q, L_q, and L_c. This graph is a directed graph, and if the term t_j is a related term of t_i, there is an edge from t_i to t_j. We apply a biased PageRank algorithm to this graph.

PageRank[3] is a method for computing the importance of Web pages using a Web link structure. The main criterion in PageRank is that a Web page is important if many other important Web pages link to it. This means that if page u has a link to page v, the link propagates the importance of u to v. Let $r(u)$ represent the degree of importance of page u, and let F_u represent the set of pages linked by page u. We can assume that all links are equal, therefore, the link (u, v) propagates $r(u)/|F_u|$ units of importance from page u to page v. Because $r(u)$ is also recursively determined by pages that point to u, the PageRank algorithm is computed using the power method. Let B_v be the set of pages that points to v, N be the number of all vertices in the graph, and α be the damping factor, then this simple idea leads to the following equation.

$$r_{i+1}(v) = \alpha \sum_{u \in B_v} \frac{r_i(u)}{|F_u|} + \frac{1 - \alpha}{N}. \qquad (4)$$

Throughout this paper, we set α to 0.85, following the original PageRank algorithm. In order to evaluate the ease of reaching each related term from q, we revise Equation 4 on the basis of biased PageRank:

$$r_{i+1}(v) = \alpha \sum_{u \in B_v} \frac{r_i(u)}{|F_u|} + (1 - \alpha)\frac{f_{ini}(v)}{\sum_{t \in C_q} f_{ini}(t)}, \qquad (5)$$

where $f_{ini}(v)$ is the initial value of vertex v and is defined by:

$$f_{ini}(v) = \begin{cases} \frac{f_{co}(v)}{\sum_{t \in C_q} f_{co}(t)} & v \in C_q \\ 0 & v \notin C_q. \end{cases}$$

Here, $f_{co}(v)$ is the score of term v calculated in Equation 3. We apply this process to the graph G_2. A vertex with a low score in L_q is a term that has a low relationship with q. However, in this step we can only consider related terms that can be directly reached through appropriate coordinate terms.

4.3.3 *Calculating the degree of relevance between a theme term and each related term (2)*

Finally, we evaluate the degree of relevance of each related term $e_i \in L_q$ by considering the coordinate terms of e_i. In the second phase, related terms such as "Indiana" do not have a high score because they are not directly related to appropriate coordinate terms of q. However, as we described, it is possible to reach "Indiana" from appropriate coordinate terms through its hypernyms such as "American state" and "place name." The objective of this phase is to increase the score of such related terms.

Given a related term $e_i \in L_q$, we first collect all of its coordinate terms and hypernyms. We denote the set of e_i and all its coordinate terms as C_{e_i} and the set of hypernyms of e_i as H_{e_i}. In C_{e_i}, some terms may be included in graph G_2, but others are not. We construct a bipartite graph that consists of C_{e_i} and H_{e_i}. Edges exist between a term $u_i \in C_{e_i}$ and a hypernym $v_j \in H_{e_i}$ when v_j is a hypernym of u_i. We apply the Co-HITS algorithm to the bipartite graph. The initial score of each hypernym is zero. If a vertex in C_{e_i} is included in graph G_2, the initial score of the vertex is the value calculated by the steps described in Section 4.3.2. If a vertex in C_{e_i} is not included in graph G_2, its initial score is zero. We calculate the score of each vertex in the following equations:

$$x_i = (1 - \lambda_u)x_i^0 + \lambda_u \sum_{v_j \in H_{e_i}} w_{ji}^{vu} y_j, \qquad (6)$$

$$y_j = (1 - \lambda_v)y_j^0 + \lambda_v \sum_{u_i \in C_{e_i}} w_{ij}^{uv} x_i, \qquad (7)$$

where x_i^0 and y_j^0 represent the initial scores for terms u_i and v_j, respectively, and x_i and y_j are the scores of u_i and v_j, respectively. Moreover, $w_{ij}^{uv} = \frac{1}{|hyper(u_i)|}$ and $w_{ji}^{vu} = \frac{1}{|hypo(v_j)|}$. In this bipartite graph, the scores of all nodes $v_j \in H_{e_i}$ are equal to 0; therefore, we set λ_v as 1. We discuss the effectiveness of the parameter λ_u in Section 5. We conduct the operation for each related term of q; let $Rel(q, e_i)$ represent the score calculated by Equation 6.

4.4 Calculating the degree of popularity of a related term

We calculate the degree of popularity of each related term by the following two methods.

The first method regards the PageRank score of articles as the popularity degree. In the PageRank algorithm, an article that is referenced by many good articles has a high PageRank score, and we assume that the title of such an article is generally well known. Hence we apply the PageRank algorithm to all articles in Wikipedia on the basis of link structure. The popularity degree of a term corresponds to the PageRank score of an article whose title is the term. We denote the PageRank score of a term e_i as $PageRank(e_i)$.

The second method considers the Web hit count of the term as the popularity degree. A term with a high hit count potentially infers frequent use of the term. We get the Web hit count of a term by

Table 1: Examples of queries (English translation).

Category	Query with more than 150 related terms	Query with fewer than 150 related terms
Person	Prince Shotoku, Tamori, Nobita Nobi	Funaki Tomosuke, Higashikuni Shigeko
Region	Monaco, The Rhine, Venus	Ohsu Domain, Kainan Island
Product	Air-bag, Train lunchi, Rocky Joe	Rhythm guitar, Two-legged robot
Facility	Nagoya Station, Theater, Tokyo Sky Tree	U.S. Library of Congress, Byodoin
Organization	UNIQLO, Japan's national soccer team, Sanyo Electric	Mitsui Group, University cooperative

using the Yahoo! Web Search API[6] and denote the hit count of e_i as $Hit(e_i)$.

4.5 Calculating the degree of unexpectedness

Given the theme term q, we can find the strength of a relationship $Rel(q, e_i)$ between q and each of its related terms e_i. We have established that there is higher degree of unexpectedness when there is a lower relationship value; therefore, we use the inverse of $Rel(q, e_i)$. For the popularity degree of each related term, a higher popularity degree results in a higher degree of unexpectedness. When we use the PageRank score, the degree of unexpectedness is calculated by the following equation:

$$Unexp(q, e_i) = \frac{1}{Rel(q, e_i)} \cdot PageRank(e_i). \qquad (8)$$

When we use the Web hit count, the degree of unexpectedness is calculated by the following equation:

$$Unexp(q, e_i) = \frac{1}{Rel(q, e_i)} \cdot log_{10} Hit(e_i). \qquad (9)$$

We use the logarithm to scale back the influence of the Web hit count because it differs significantly from one term to another.

5. EXPERIMENT

We conducted an experiment to examine the effectiveness of our proposed method. The objective of our experiment is to clarify two research questions: (1) Is considering the degree of popularity of related terms important to the discovery of unexpected information? (2) Is considering the relationship between coordinate terms of a theme term and coordinate terms of its related terms important to the discovery of unexpected information?

To answer these questions, we used six proposed methods and compared them with four simpler methods. The six proposed methods calculate the degree of unexpectedness of each related term by using Equation 8 or 9. In order to compare the impact of λ_u in Equation 6, we set λ_u to 0.25, 0.5, and 0.75. A method using Equation 8 in which λ_u was set to 0.25 was denoted as PR_{25}. Similarly, we denote the other methods as PR_{50} and PR_{75}. We denote methods using Equation 9 with λ_u set to 0.25, 0.5, and 0.75 as HIT_{25}, HIT_{50}, and HIT_{75}, respectively.

We use three additional simple methods to answer the first research question. In these methods, only the strength of the relationship between a theme term and a related term is evaluated and the popularity degree of related terms is neglected. The score of related term e_i of the theme term q is calculated by $Unexp(q, e_i) = \frac{1}{Rel(q, e_i)}$. In these methods, we also set λ_u to 0.25, 0.5, and 0.75. We denote each method as REL_{25}, REL_{50}, and REL_{75}.

We also proposed a simple method to answer the second research question. In this method, we get the Web hit count of each pair of (q, e_i) using the Yahoo! Web Search API. The query is "q $\wedge e_i$" for the pair of (q, e_i), then the related terms are ranked in ascending order. That is, we assume that if a related term has low co-occurrence frequency with q, the term is unexpected for q. We denote this method as TC.

We discover unexpected information relative to a theme term from a Wikipedia article where the title of an article is the theme term. Given a theme term and a related term, we extract a sentence that includes the related term from the article. If the related term is included in more than one sentence, we extract the first sentence that uses the term. In the experiment, described in Section 5.2, this sentence will be used as the corresponding information.

5.1 Query set

We created a query set that consisted of 75 theme terms in five categories: names of people, facilities, regions, products, and organizations. Each category included 15 theme terms. Obviously, if a user is not at all familiar with the theme term, all information will be not unexpected for him. Hence we selected terms that appeared in the top 5% of PageRank scores among all Wikipedia articles. The number of articles was $17,325$. Moreover, we assumed that the fewer the number of related terms, the lower would be the probability of discovering unexpected information. To examine this, we first divided the set of articles into two groups: group (a) included articles that had more than 150 related terms, and group (b) included articles that had less than 150 related terms. There were $4,854$ articles in group (a) and $12,471$ articles in group (b). We randomly selected 10 articles for each category from group (a). The remaining five terms in each category were randomly selected from group(b). We used the title of each article as a query, or a theme term. Examples of the query set are shown in Table 1.

5.2 Procedure

In this experiment, we recruited five evaluators and administered a questionnaire. Two males and a female were in their thirties and two females were in their twenties.

We created the questionnaire as follows. First, a theme term was used with each method. Given a theme term, each of the ten methods returned a ranked list of related terms in descending order by the degree of unexpectedness. We used the top five related terms from each method. We pooled the related terms and generated a list of randomly sorted pairs of related terms and the corresponding information. We asked evaluators to label each pair of a related term and its information on a scale of 1-4 from expected to unexpected by asking "Do you think this information is unexpected?"

A total of 75 questionnaires were constructed, with each questionnaire corresponding to a single theme term. We then ordered five sets of questionnaires taking the order effect into consideration. Five evaluators answered the questionnaires individually. Then we calculated the average degree of unexpectedness for each piece of information. For example, for a query "Monaco," one method detected the related term "Kimiko Date[7]" as a highly unexpected term

[6] http://developer.yahoo.co.jp/webapi/search/websearch/v1/websearch.html

[7] Kimiko Date is a famous Japanese tennis player and is also a television personality.

Table 2: Performance comparison of each category for ten methods measured by nDCG

Method	Person	Region	Product	Facility	Organization	Average
TC	0.705	0.757	0.773	0.787	0.780	0.760
REL$_{25}$	0.805	0.792	0.837	0.800	0.853	0.817
REL$_{50}$	0.807	0.803	0.839	0.800	0.857	0.821
REL$_{75}$	0.807	0.808	0.841	0.804	0.852	0.822
PR$_{25}$	**0.828**	0.830	0.846	**0.825**	0.860	**0.838**
PR$_{50}$	0.824	0.830	0.851	0.821	0.860	0.837
PR$_{75}$	0.818	0.836	**0.858**	0.820	0.854	0.837
HIT$_{25}$	0.798	0.832	0.843	0.824	0.856	0.830
HIT$_{50}$	0.791	0.834	0.843	0.823	0.867	0.832
HIT$_{75}$	0.790	**0.838**	0.847	0.822	**0.872**	0.834

Table 3: Performance comparison of each category for ten methods measured by NWRR

Method	Person	Region	Product	Facility	Organization	Average
TC	0.307	0	**0.478**	0	0	0.157
REL$_{25}$	**0.513**	0.118	0.319	0.215	0.165	0.266
REL$_{50}$	0.506	0.118	0.327	0.199	0.177	0.266
REL$_{75}$	0.506	0.163	0.332	0.194	0.177	0.274
PR$_{25}$	0.434	0.184	0.341	0.418	0.194	0.314
PR$_{50}$	0.418	0.184	0.361	0.241	0.194	0.280
PR$_{75}$	0.421	0.199	0.361	0.241	0.194	0.283
HIT$_{25}$	0.384	0.165	0.234	**0.567**	0.118	0.293
HIT$_{50}$	0.384	0.184	0.234	0.542	0.172	0.303
HIT$_{75}$	0.400	**0.218**	0.234	0.513	**0.241**	**0.321**

and output the corresponding information "Kimiko Date is now living in Monaco." The five evaluators assessed the unexpectedness of this information as 4, 3, 2, 3, and 2. The average degree of unexpectedness of this information was 2.8.

5.3 Metrics for evaluation

We used Normalized Discounted Cumulated Gain (nDCG)[10] and Normalized Weighted Reciprocal Rank(NWRR)[19] as evaluation metrics.

nDCG is a measure of retrieval effectiveness that utilizes graded relevance judgments. We used this metric because it is preferable to rank much and more unexpected information at higher rank.

NWRR only considers a correct answer at the highest rank. In our experiment, each piece of information has an unexpected score judged by evaluators. This score ranges from 1 to 4, and we regard information with a score larger than the intermediate value of 2.5 as especially unexpected information, or an answer. Due to lack of space, we refer the reader to a paper by Sakai [19] for further detail of this metric. In this metric, the score is equal to 1 if a method can put the most unexpected information for a theme term at rank 1. If the top five pieces of information discovered by a method are judged not unexpected by evaluators, the NWRR score is 0.

5.4 Experimental results

The nDCG scores for each method and category are shown in Table 2. In all categories, one of the six proposed methods that considered the popularity degree of related terms resulted in the highest nDCG. REL_{25}, REL_{50}, and REL_{75} followed those six methods. The results show that it is important to consider the popularity degree of related terms to discover unexpected information. The TC method got the lowest scores in all categories. This result shows the importance of considering the relationship between coordinate terms of both a theme term and its related terms to discover unexpected information. In this experiment, there was not a significant difference between the methods that used the PageRank score and methods that used the Web hit count for calculating the popularity degree of related terms. As for the parameter λ_u in the Co-HITS algorithm, too, there was not a significant difference. The NWRR scores for each method in each category are shown in Table 3. On an average, HIT_{75} could discover more unexpected information at a higher rank than other methods. TC and REL_{25} got the highest scores in the categories of product and person, respectively. However, the average scores of these two methods were lower than our six proposed methods and the average nDCG scores were also lower. These results indicate that we could discover unexpected information by chance even if we did not consider the degree of popularity and the relationships between terms. On the other hand, our proposed methods could discover unexpected information in any category.

We show some examples of information evaluated as unexpected information in Table 4. For the theme term "Akita Prefecture," an unexpected related term "lifestyle-related disease" and the corresponding information "In addition to excessive drinking, people consume too much salt from preserved foods such as pickles and Akita Prefecture has a high rate of death from lifestyle-related disease such as a stroke." was discovered. In our method, other prefectures in Japan were evaluated as appropriate coordinate terms of "Akita Prefecture," and disease names were evaluated as appropriate coordinate terms of "lifestyle-related disease." In general, a prefecture does not have a relationship with a specific disease, and "lifestyle-related disease" is a well-known term. Hence, our method could evaluate the related term as an unexpected term.

Finally, in Table 5, we show the number and ratio of theme terms in which we could discover at least one piece of unexpected information. On an average, we could discover unexpected information in 40% of theme terms that had greater than 150 related terms and in 24% of theme terms that had less than 150 related terms. This result shows that the probability of discovering unexpected information is high if a theme term has many related terms. According to our observations, there are two principal reasons why our methods could not discover unexpected information. One reason is that unexpected information is not written in some articles even when the theme term has many related terms. This tendency was especially true in the building, facility, and organization categories. The other reason stems from specific characteristics of our method. For example, the article for "digital camera" includes the information "A digital camera is often abbreviated to Dejikame in Japan, but Dejikame is a registered trademark of SANYO Electric and other companies" and this information seems to be unexpected. The related term in this information is "SANYO Electric," but it is related to many other electrical products that are appropriate coordinate terms of "digital camera." Therefore, our method could not discover this information.

6. CONCLUSION

In this paper, we proposed a new method for discovery of unexpected information. In particular, we focused on two aspects: (i) the relationship between a theme term, its coordinate terms, its related terms, and their coordinate terms, and (ii) the degree of popularity of each related term. We conducted an experiment to clarify the importance of considering these two aspects. Our results showed that the degree of popularity of a related term was highly relevant to the degree of unexpectedness. Moreover, it was also effective to consider the coordinate terms rather than considering only the co-occurrence frequency of a theme term and its related term.

We would like to explore methods for determining unexpected information from other information resources. This would enable us to find a variety of unexpected information; however, we would

Table 4: Examples of discovered unexpected information (English translation).

Theme term	Related term	Unexpected information
Air-bag	Fire Defense Law	The air-bag was never developed in Japan because using gunpowder was prohibited by the Fire Defense Law at the time.
Horyuji temple	Cultural Property Fire Prevention Day	The Law for the Protection of Cultural Properties was established because of a fire disaster, and in response, the government designated January 26 as Cultural Property Fire Prevention Day.
Vending machine	scenery	A light pollution problem and its disadvantageous effect on scenery are pointed out.
Monaco	Kimiko Date	Kimiko Date is now living in Monaco.
Mitsui Group	Tokyo Disneyland	Sumitomo Mitsui Banking has branches in Tokyo Disneyland and Tokyo Disney SEA.
Akita Prefecture	lifestyle-related disease	In addition to excessive drinking, people consume too much salt from preserved foods such as pickles, and Akita Prefecture has a high rate of death from lifestyle-related diseases such as a stroke.
Train lunch	earthen teapot	In 1992, the Japanese Railway Ministry banned the use of earthenware teapots for hygienic reasons; glass teapots were introduced.
Akira Toriyama	Fabre	Akira Toriyama designed the cover and frontispiece of "The Insect World of J. Henri Fabre" that was edited and translated by Daisaburo Okumoto and published by Shueisha.
Nobita Nobi[8]	first-degree equation	He solved a difficult first-degree equation "$3/8x = 9/10$" and got a score of 100.

need to address the problem of removing noise terms. In addition, we need to consider the credibility of unexpected information especially when we discover unexpected information from more general Web pages. False or untrue information is not useful. One method to verify credibility is to check the publisher of the information. If the unexpected information has been written by an expert in the domain, it is more likely that the information is credible. We intend to undertake this work in the future.

Table 5: The number and the ratio of theme terms that could find unexpected information.

Category	Over 150	Under 150	Total
Person	60% (6/10)	60% (3/5)	60% (9/15)
Region	30% (3/10)	0% (0/5)	20% (3/15)
Product	50% (5/10)	40% (2/5)	47% (7/15)
Facility	40% (4/10)	0% (0/5)	27% (4/15)
Organization	20% (2/10)	20% (1/5)	20% (3/15)

Acknowledgements

This work was supported in part by the following projects: Grants-in-Aid for Scientific Research (Nos. 24240013, 24680008) from MEXT of Japan and JSPS KAKENHI Grant Number 243993.

7. REFERENCES

[1] E. Agichtein and L. Gravano. Snowball: extracting relations from large plain-text collections. In *Proc. of ACM DL 2000*, pages 85–94, 2000.

[2] G. Berger and A. Tuzhilin. Discovering unexpected patterns in temporal data using temporal logic. In *Temporal Databases Research and Practice, Lecture Notes in Computer Science 1399*, pages 281–309, 1998.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. of WWW 2007*, pages 107–117, 1998.

[4] A. Carlson, J. Betteridge, R. C. Wang, E. R. Hruschka, Jr., and T. M. Mitchell. Coupled semi-supervised learning for information extraction. In *Proc. of ACM WSDM 2010*, pages 101–110, 2010.

[5] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. Learning to construct knowledge bases from the world wide web. *Artif. Intell.*, 118(1-2):69–113, 2000.

[6] H. Deng, M. R. Lyu, and I. King. A generalized Co-HITS algorithm and its application to bipartite graphs. In *Proc. of ACM SIGKDD 2009*, pages 239–248, 2009.

[7] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates. Web-scale information extraction in knowitall: (preliminary results). In *Proc. of WWW 2004*, pages 100–110, 2004.

[8] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *Proc. of VLDB 2004*, pages 576–587, 2004.

[9] T. H. Haveliwala. Topic-sensitive pagerank. In *Proc. of WWW 2002*, pages 517–526, 2002.

[10] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

[11] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46:604–632, September 1999.

[12] B. Liu, Y. Ma, and P. S. Yu. Discovering unexpected information from your competitors' web sites. In *Proc. of ACM SIGKDD 2001*, pages 144–153, 2001.

[13] A. Nadamoto, E. Aramaki, T. Abekawa, and Y. Murakami. Searching for important but neglected content from community-type-content. In *Proc. of SITIS 2008*, pages 161–168, 2008.

[14] A. Nadamoto, E. Aramaki, T. Abekawa, and Y. Murakami. Content hole search in community-type content. In *Proc. of WWW 2009*, pages 1223–1224, 2009.

[15] Y. Noda, Y. Kiyota, and H. Nakagawa. Discovering serendipitous information from wikipedia by using its network structure. In *Proc. of ICWSM 2010*, pages 299–302, 2010.

[16] B. Padmanabhan and A. Tuzhilin. A belief-driven method for discovering unexpected patterns. In *Proc. of ACM SIGKDD 1998*, pages 94–100, 1998.

[17] B. Padmanabhan and A. Tuzhilin. Small is beautiful: discovering the minimal set of unexpected patterns. In *Proc. of ACM SIGKDD 2000*, pages 54–63, 2000.

[18] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proc. of ACM SIGIR 1994*, pages 232–241, 1994.

[19] T. Sakai. On the properties of evaluation metrics for finding one highly relevant document. *Information and Media Technologies*, 2(4):1163–1180, 2007.

[20] K. M. Svore and C. J. Burges. A machine learning approach for improved BM25 retrieval. In *Proc. of ACM CIKM 2009*, pages 1811–1814, 2009.

[8] Nobita Nobi is a famous cartoon character who is not a good student.