Title	A Study on Web Search and Analysis based on Typicality(Dissertation_全文)
Author(s)	Tsukuda, Kosetsu
Citation	Kyoto University (京都大学)
Issue Date	2014-09-24
URL	http://dx.doi.org/10.14989/doctor.k18617
Right	
Type	Thesis or Dissertation
Textversion	ETD

A Study on Web Search and Analysis based on Typicality

Kosetsu Tsukuda

ABSTRACT

With the increase in the amount of information on the Web, the number of people who access web-based information increases. A general and common goal of a web search is to find information about an unknown topic. Although there are various ways to support users when they search for a topic, one solution is to realize a search based on typicality. Cognitive psychology suggests that showing typical instances in a category is useful to understand the outline of the category. After understanding the outline of the category, it is helpful to achieve greater understanding of the category by showing atypical examples and unexpected examples. Therefore, in this thesis, we focus on searches for and analysis of data based on typicality and unexpectedness. We consider the two types of typicality of information, such as an object, an object set, and a relation. The two types are "typicality based on data (TD)" and "typicality based on social recognizability (TSR)." We also consider typicality based on central tendency and frequency of instantiation, which were proposed in cognitive psychology, for TD and TSR. This thesis includes the following three research topics:

1. **Search for an Object Set based on Typicality**
   We propose a method for calculating the typicality of an object set (e.g., a recipe and a tourist route) that consists of some objects (e.g., ingredients and tourist spots). First, we compute the typicality of an object set based on our own hypothesis of typicality. The typicality is calculated based on the appearance frequency of each object and the co-occurrence frequencies between objects. We also propose methods for recommending candidate objects for addition to and deletion from an object set to change it to a more typical or atypical set. In addition, we focus on two viewpoints of typicality (i.e., central tendency and frequency of instantiation) that were proposed in cognitive psychology using recipes as the target object set. We compare the typicality of a recipe judged by assessors with that calculated from each viewpoint.

2. **Discovering Unexpected Information based on the Popularity of Terms and the Typicality of Relationships between Terms**

We propose a method for discovering unexpected information for a given query. Given a query q (e.g., "Hiromitsu Ochiai"), our method first detects an unexpected related term e (e.g., "Gundam") and then presents unexpected information (e.g., "Hiromitsu Ochiai is a Gundam maniac."). We hypothesize that information is unexpected when it includes a related term that has an atypical relationship with the query in TD and the popularity of the related term is high. Based on this hypothesis, we compute unexpectedness by considering the relationships between coordinate terms of q and coordinate terms of e, and e's popularity. Experimental results show that considering these two factors are effective for discovering unexpected information.

3. **Measuring Perceived Strength of the Relationship between Terms to Discover an Unexpected Relationship**

   The strength of the relationship between terms in TD is not necessarily correspond to that in TSR. We hypothesize that when the strength of the relationship between the terms is high (low) in TD but low (high) in TSR, the relationship is unexpected. Several methods have been proposed to compute the strength of the relationship between terms based on Wikipedia data or co-occurrence frequencies of the terms on the Web. These methods reflect the strength of the relationship in TD. We propose a method for computing the perceived strength of relation between terms (an attribute and an object). The proposed method considers two factors: (1) the popularity of an object, and (2) the strenght of the relations between an attribute and an object's coordinate terms. We utilize crowdsourcing to collect data of the perceived strength of a relation between an attribute and an object, and evaluate the proposed method. We also verify the aforementioned hypothesis using a crowdsourcing.

ACKNOWLEDGEMENT

I would like to express my deep gratitude to my supervisor Professor Katsumi Tanaka for his valuable advice and comments for six years. I have great estimation for his research visions and ability of abstraction.

I am grateful to my thesis committee members Professor Masatoshi Yoshikawa and Professor Sadao Kurohashi for their helpful comments and suggestions about this thesis.

I would like to thank my research advisor Professor Hayato Yamana at Waseda University for his valuable discussions during my PhD course. I have learnt many important things about research from him, which can be found in many parts of my current researches.

I would like to thank Associate Professor Hiroaki Ohshima, Assistant Professor Takehiro Yamamoto, and Assistant Professor Makoto P. Kato for many comments about my researches. It was impossible to write up this thesis without their supports.

I would like to show my great appreciation to Associate Professor Satoshi Nakamura at Meiji University. He told me presentation techniques and the fun of programming. I would not be who I am without him.

I am grateful to Associate Professor Tetsuya Sakai at Waseda University for his valuable discussions and suggestions in Microsoft Research Asia.

I wish to thank Professor Osami Kagawa at Osaka Gakuin University, Associate Professor Yoko Yamakata, and Associate Professor Adam Jatowt for their good comments during the laboratory meeting.

I want to thank secretaries of Professor Tanaka: Ms. Ikebe, Ms. Sato, and Ms. Shiraishi. I also want to thank my colleagues in Tanaka laboratory for their cooperation and fruitful discussions.

Finally, I want to say thank you to my parents Takaki Tsukuda and Kaori Tsukuda for their strong support.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Background

With the increase in the amount of information on the Web, the number of people who access Web-based information increases. There are novice and expert users among those who search for information; some people have sufficient knowledge of the target domain to conduct an effective search and others do not. In addition, the goal of a Web search varies from person to person; for example, "I want to read a Web page that explains about a topic," "I want to buy something," "I want to know the latest information about a topic," "I want to know the reputation about something," etc.

As more people access the Web for various reasons, various types of search methods have been proposed. From the viewpoint of the dimensions of relevance, methods based on the similarity between a query and a document were proposed initially [59, 63], and then methods based on link analysis were proposed [7, 15, 16, 28]. Methods based on diversity were proposed [1, 20, 68] to address the fact that novice users often input short [17] and ambiguous [77] queries. In addition, methods based on freshness [14, 19] and novelty [11] have also been proposed. Recently, the dimensions of relevance based on cognitive perspectives (i.e., how users feel when they see retrieved information) have attracted increasing attention. For example, Fox *et al.* [23] and Hassan and White [27] estimated whether a user was satisfied with each web page in the search result of a commercial search engine and proposed retrieval methods based on the degree of satisfaction. Kato *et al.* [34] proposed the concept of cognitive search intents (CSIs). They focused on exhaustiveness, comprehensibility [2, 52], subjectivity and objectivity [80], and concreteness and abstractness [74], and administered a questionnaire-based user study. They reported that over 50% of the subjects occasionally had experience with searches with CSIs, and approximately half

of the subjects did not input any keywords representing CSIs.

A general and common goal of a Web search is to find information about an unknown topic. Broder [8] classified Web queries into three classes according to their intent, i.e., navigational (the immediate intent is to reach a particular site), informational (the intent is to acquire some information assumed to be present on one or more Web pages), and transactional (the intent is to perform some Web-mediated activity). He reported that approximately half of the queries logged by AltaVista log could be classified as informational. Rose and Levinson [65] also reported that approximately 60% of the queries had informational search intent, which was represented by "my goal is to learn something by reading or viewing Web pages." In the questionnaire conducted by Nakamura *et al.* [51], approximately 83% of the users responded that the most significant reason for their Web search was to obtain information about particular things. Although there are various ways to support users when they search for a topic, one solution is to realize a search based on typicality. Cognitive psychology suggests that showing typical instances in a category is useful to determine the outline of the category [43]. After understanding the outline of the category, it is helpful to achieve greater understanding of the category by showing atypical examples and unexpected examples. Typicality and unexpectedness are the classes of CSIs. Hence, it is difficult for users to input appropriate queries to search for typical or atypical information.

There are several possible problems with searches based on typicality and unexpectedness. These problems can be summarized as follows:

- For example, when a user wants to search a recipe for typical pasta carbonara or search unexpected information about Kyoto, it is not effective to input queries such as "carbonara typical" and "Kyoto unexpected" because the keyword "typical" or "unexpected" is not always included in a Web page that contains typical or unexpected information.

- Even if typical (atypical) information is provided in a web page, a user cannot judge whether the information is truly typical (atypical) when he does not have sufficient knowledge about the domain.

- It is difficult to find atypical and useful information or unexpected information because considerable noisy information is included in atypical information about a topic.

Therefore, we focus on searches for and analysis of data based on typicality and unexpectedness. We propose search methods and evaluate their effectiveness.

1.2 Approach

Before developing search methods or analyzing information based on typicality, it is necessary to define typicality. In the field of cognitive psychology, many studies that focus on typicality

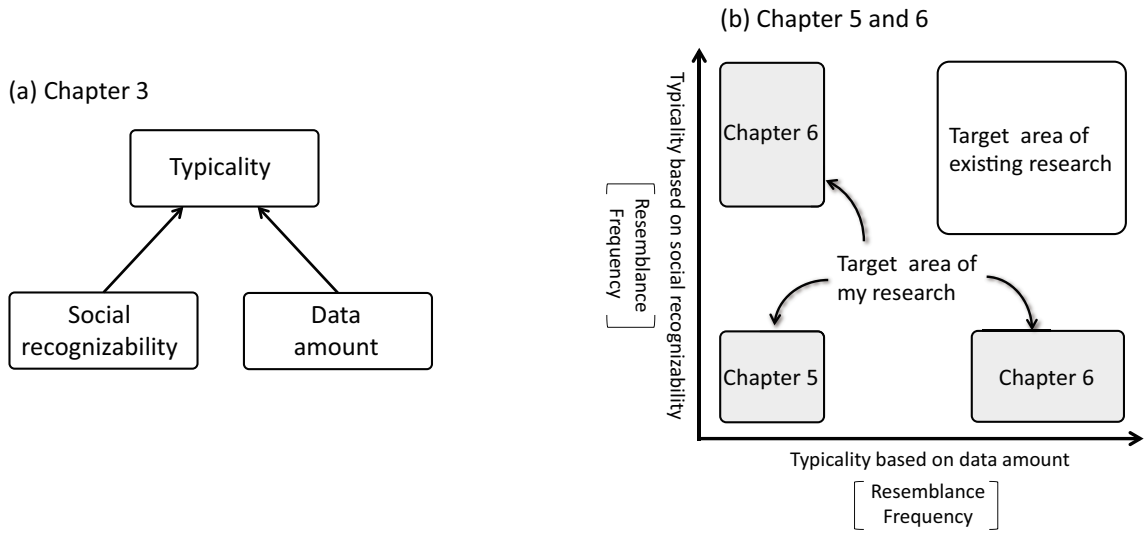

Figure 1.1: Structure of the thesis and our research position.

have been conducted. Some major concepts of typicality discussed in these studies are *central tendency* (CT), *frequency of instantiation* (FOI), and *ideals* [4]. In the CT concept, the more an object is similar to other objects in the same category, the more typical it is. In the FOI concept, the more often one has experienced an object in a category, the more typical the object is. In the concept related to ideals, the more an object is similar to a goal associated with its category, the more typical it is. The detail of each concept will be described in Section 2.2.1. Based on these ideas, we consider resemblance-based typicality and frequency-based typicality.

In addition, we consider the two types of typicality of information, such as an object, an object set, and a relation. The two types are "typicality based on data amount (TD)" and "typicality based on social recognizability (TSR)." TD represents typicality in an information source (e.g., the Web). TSR represents typicality that reflects people's cognition to the object. In information retrieval and data mining studies, researchers have generally assumed that information with high TD has high TSR and have proposed methods for discovering information with high TD. However, TD does not necessarily correspond to TSR. There is significant amount of information that have (1) high TD and low TSR, (2) low TD and high TSR, and (3) low TD and low TSR. We target such information and search and analyze information that is not typical but useful, as discussed in Section 1.1.

Typicality as treated in this research is shown in Figure 1.1. We also consider typicality based on resemblance and frequency for TD and TSR. When we compute TD, the typicality of an object (as well as an object set and a relation between terms) based on resemblance is represented by "how many objects similar to the object exist in an information source," and typicality based on

frequency is represented by "how frequently the object appears in an information source." When we compute TSR, typicality based on resemblance is represented by "people think the object is typical because they have experienced many objects similar to the object," and typicality based on frequency is represented by "people think the object is typical because they often see it." Based on these ideas, we first target an object set and estimate typicality based on TD and TSR (Figure 1.1(a)). The effectiveness of each viewpoint is evaluated by comparing the typicality computed from each viewpoint with that judged by assessors (Chapter 3). Next, we discover information that has (1) high TD and low TSR, (2) low TD and high TSR, and (3) low TD and low TSR (Figure 1.1(b)). Specifically, we tackle the following two research topics.

- First, we target a relationship between terms and analyze a relationship that has (3) low TD and low TSR. We compute TD based on resemblance. Some relationships in (3) are useful or unexpected, but others are not. To distinguish them, we consider the recognizability of a term. Given a query, we detect a term that has an unexpected relationship with the query and discover unexpected information that includes the term (Chapter 5).

- Second, we target a relationship between terms and analyze a relationship that has (1) high TD and low TSR or (2) low TD and high TSR. TSR is computed based on both resemblance and frequency, and TD is computed based on frequency. We verify whether a relationship in (1) and (2) is unexpected and estimate the unexpectedness of a given term pair (Chapter 6).

1.3 Thesis Organization

The remainder of this thesis is organized as follows:

- Chapter 2
  This chapter describes the study related to the research presented in this thesis.

- Chapter 3
  This chapter proposes a method for calculating the typicality of an object set such as a recipe and a tourist route. An object set consists of some objects such as ingredients and tourist spots. First, we compute typicality based on our own hypothesis of typicality. The proposed method first detects the most typical set of objects in a category based on the appearance frequency of each object and the co-occurrence frequencies between objects. Given an object set, we compute its typicality based on the affinity between its objects and the difference between the object set and the most typical set of objects. In addition, we

propose methods for recommending candidate objects for addition to and deletion from an object set to change it into a more typical or atypical set. Finally, we focus on two viewpoints of typicality, i.e., central tendency and frequency of instantiation, as proposed in cognitive psychology. We evaluated the effectiveness of each viewpoint for estimating typicality of an object set by using recipes as the target object set.

- Chapter 4

  In this chapter, methods for ranking coordinate terms and hypernyms of a given query according to their appropriateness are proposed. Although previous studies have proposed methods for discovering coordinate terms or hypernyms of a query, they focused on only discovering such terms and evaluating discovered terms based on a binary evaluation: appropriate or inappropriate. Unlike these studies, we rank coordinate terms and hypernyms of a query and evaluate the terms by considering their appropriateness. In the proposed method, a bipartite graph is created based on hypernyms of a query and hyponyms of each hypernym using a hypernym-hyponym dictionary. Subsequently, we apply a HITS-based algorithm to the bipartite graph and rank coordinate terms and hypernyms based on their appropriateness. The experimental results obtained using 50 queries demonstrate that our method could rank appropriate coordinate terms and hypernyms higher than other comparable methods. Methods proposed in this chapter are used in Chapter 5 and 6 to rank coordinate terms for a given query.

- Chapter 5

  This chapter proposes a method for discovering unexpected information for a given query. For example, given a query "Hiromitsu Ochiai," our proposed method discovers unexpected information "Hiromitsu Ochiai is a Gundam maniac." In this chapter, we target information that contains two objects, i.e., a query keyword and its related term. In the above example, "Gundam" is the related term of "Hiromitsu Ochiai." We hypothesize that information is unexpected when it includes a related term that has an atypical relationship with the query in TD and the popularity of the related term is high. We compute the typicality of the relationship between a query and its related term based on the relationships between the coordinate terms of the query and those of its related term.

- Chapter 6

  In this chapter, we focus on the difference between the strength of the relationship between terms in TD and that in TSR. We hypothesize that when the strength of the relationship between the terms is high (low) in TD but low (high) in TSR, the relationship is unexpected. To verify this hypothesis, we propose a method for computing the perceived strength of the

relationship between the terms (an attribute and an object). The proposed method considers two factors: (1) the popularity of an object and (2) the strength of the relationships between an attribute and an object's coordinate terms. We conduct experiments using 25 attributes that were included in five categories: country, vegetable, tourist spot in Kyoto, electronic company, and baseball player. We utilize crowdsourcing to collect data of the perceived strength of a relation between an attribute and an object, and evaluate the proposed method. We also verify the aforementioned hypothesis using a crowdsourcing.

- Chapter 7
  This chapter summarizes this thesis and addresses some directions to be explored in the future study.

CHAPTER 2

RELATED WORK

2.1 Dimensional Relevance

The concept of relevance has been studied in information retrieval. One of the most primitive concepts is the similarity between a query and a document. Similarity has been computed by weighting search terms [63] or by applying a probabilistic language model [59]. In addition, link analysis methods have been proposed [7, 15, 16, 28], where it was assumed that a document that was linked by many important documents was important.

Concepts beyond topical relevance have also been proposed. One such concept is diversity-oriented search [1, 20, 68]. Search result diversification is necessary for novice web search users because they often input short [17] and ambiguous [77] queries. Another proposed concept is freshness [14, 19], which considers the timeliness of a web document.

Recently, search measurements based on cognitive viewpoints (i.e., how users feel when they see retrieved information) have attracted increasing attention. For example, Fox *et al.* [23] and Hassan and White [27] estimated whether a user was satisfied with each web page in the search result of a commercial search engine and proposed retrieval methods based on the degree of satisfaction. These studies indicate the limitation of information retrieval simply based on the relevance and the popularity of web pages. Other dimensions of relevance that reflect cognitive search intents also have been proposed. Akamatsu *et al.* [2] and Nakatani, Jatowt, and Tanaka [52] proposed the concept of comprehension-based web searches. Nakatani, Jatowt, and Tanaka measured the comprehensibility of web pages by considering both document readability and the difficulty proposed by technical terms in search queries based on Wikipedia link analysis. Yu and Hatzivassiloglou [80] proposed a method for separating opinions from fact at both the document and the sentence level. This method enables users to retrieve subjective web pages.

Tanaka *et al.* [74] proposed a method for computing the concreteness of documents by aggregating the predicted concreteness of terms. Kato *et al.* [34] investigated query formulations by users with cognitive search intents (CSIs) CSIs represent user requirements for the cognitive characteristics of documents to be retrieved. Kato *et al.* focused on exhaustiveness, comprehensibility [2, 52], subjectivity and objectivity [80], and concreteness and abstractness [74], and administered a questionnaire-based user study. They reported that over 50% of subjects occasionally had experiences with searches that involved CSIs and that approximately half of the subjects did not input any keywords representing CSIs.

2.2 Typicality

2.2.1 Typicality in Cognitive Psychology

The members of a category differ in the extent to which they are a good example of the category [4]. In cognitive psychology, the degree of goodness of an object is regarded as its typicality in the category [48]. That is, an object has different degrees of typicality in different categories. For example, an object "dog" has different degree of typicality in the categories "mammal" and "meat source." The prototype theory [41] is an early typicality theory wherein a category is represented by a best prototype. The category prototype consists of all the salient properties of the objects that are classified into the category [79]. An object is considered more typical of a category, the more similar it is to the prototype.

Barsalou [4] surveyed the relationship between the typicality judged by subjects and that measured by three characteristics of typicality: *central tendency* (CT), *frequency of instantiation* (FOI), and *ideals* (I). In the concept of *CT*, the more an object is similar to other objects in the same category, the more typical it is. For example, "dog" is very similar to other members of the category "mammals," but "whale" is not as similar. Consequently, "dog" is more typical of "mammals" than "whale." The prototype theory is a type of CT, and it is known that similarity to a prototype and similarity to other objects are functionally equivalent at the level of predicting typicality [3]. In the concept of I, the more an object is similar to a goal associated with its category, the more typical it is. For example, in the category "foods to eat on a diet," the ideal is "zero calories." Therefore, people judge "agar" as more typical than "pizza" in the category. Most categories have more than one ideal. In the category "foods to eat on a diet," "it is digested slowly" is also an ideal. In the concept of FOI, the more often one has experienced an object in a category, the more typical the object is. For example, "Kinkakuji" is often introduced on TV and other media as a sightseeing spot in Kyoto; consequently, people often visit it. Therefore, people judge "Kinkakuji" as typical of the category "sightseeing spots in Kyoto." The experiments conducted by Barsalou [4] show that the characteristic of typicality that correlates strongly with

the typicality judged by subjects varies from one category to the other. In some categories, there are characteristics that have a high correlation with the typicality judged by subjects. This means that the degree of typicality is judged based on various characteristics of typicality.

2.2.2 Typicality in Computer Science

Some studies have proposed methods to calculate the typicality of an object. Rifqi [61] proposed a method to build fuzzy prototypes for fuzzy data from large databases based on the prototype theory. Lesot, Mouillet, and Bouchon-Meunier [38] adapted the method to crisp data and proposed methods to compute object typicality. In their methods, object typicality was computed based on both within-class resemblance and dissimilarity to other classes. Yeung and Leung [79] defined an object on ontology as a property vector, and proposed a method to calculate object typicality based on the prototype theory. They assumed the existence of sub-concepts in a concept. For example, sub-concepts of the concept "bird" are "sparrow," "parrot," "robin," and so on. They constructed a prototype in a concept by aggregating properties of its sub-concepts and computed object typicality. Although these studies relied on the prototype theory, some methods, such as TextRank [44] and VisualRank [33], are appropriate for computing object typicality based on CT. In these methods, objects (e.g., documents and images) are connected by edges. The weight of an edge is computed based on the similarity between the objects; the objects that have many similar objects have high scores. Cai and Leung [9] proposed a method for calculating the typicality of an object based on CT and FOI. They defined a prototype salience vector to indicate the FOI of each group of similar instances. In contrast, in the study reported here, we compute the FOI based on the appearance of frequency of each object. Cai *et al.* [10] also proposed a method for recommending objects to users based on the typicality of the user.

The related study described in this section only considers object typicality, while we consider the typicality of relationships. In the case of a relation between terms, the properties of the relationship are not noticeable; therefore, we need to develop a method to compute relationship typicality without relying on properties. Moreover, although previous related study focused exclusively on detecting typical objects, we consider atypical objects and analyze objects and term relationships by combining multiple characteristics of typicality.

2.3 Unexpectedness

2.3.1 Unexpectedness in Recommendations

In the field of information recommendation, initial recommendation systems emphasized recommendation accuracy [60]. More recently, many studies have placed importance on unexpectedness and serendipity. The unexpectedness of a recommendation list is computed based on the

difference between a set of recommendations generated by a primitive prediction model and that generated by a proposed recommender system [25]. Specifically, a set of unexpected recommendations is defined by $UNEXP = RS \setminus PM$, where RS is a set of recommendations generated by a primitive prediction model and *PM* is recommendations generated by a proposed recommender system. Serendipity is defined as a measure of the extent to which the recommended items are both attractive and surprising to the users [30]. Based on this definition, Ge *et al.* [25] defined serendipitous recommendations as recommendations that are both unexpected and useful. To achieve serendipity-oriented recommendations, a method that diversifies items in a recommendation list [82] and a method that presents items that have low similarity to a user's profile [31] were proposed. Oku and Hattori [56] designed a system that recommends serendipitous items by mixing features of two user-input items. In the field of information recommendation, unexpectedness is computed mechanically, while in this research, unexpectedness is evalulated by querying assessors. This approach allows us to discuss unexpectedness in greater depth.

2.3.2 Unexpectedness in Web Searches

To the best of our knowledge, very few studies have focused on discovering unexpected information [39, 42, 50, 54]. Noda *et al.*[54] 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 was a premier of Japan 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.*[50] 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 [49]. Liu et al.[39] proposed methods to help a company find unexpected information from competitors' Web sites by comparing their Web sites with those 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, while our objective is to discover unexpected information for a keyword. Majova *et al.* [42] proposed a method to discover unexpected information for an input query. They assumed that a term is unexpected for a query if the term appears infrequently in a document set and has high co-occurrence frequency with the query. In terms of "the degree of typicality of a relation" and "the popularity of a term," they focus on terms that have a typical relationship with a query and a low popularity. In contrast, we focus terms that have an atypical relationship with a query and a high popularity.

CHAPTER 3

SEARCH FOR AN OBJECT SET BASED ON TYPICALITY

3.1 Introduction

Internet users can now find a great variety of information on the web where the amount of information and the number of Web services have been increasing rapidly. In order to search information more efficiently, users often want to browse search results from a certain viewpoint, such as degree of freshness, credibility, specialty, and so on. Furthermore, there are many situations in which users might want to search based on the degree of typicality of information, as shown in the following example.

- A user wants to cook pasta carbonara, and searches a recipe. It is his/her first time cooking; therefore, he/she plans to cook a typical version of the dish and is in search of a supporting recipe.

- A user plans to travel to Kyoto and searches for a tourist route. He/she has already visited some typical sightseeing spots in Kyoto; therefore, the user wants to find a tourist route composed of atypical spots.

- A user plans to begin studying about Ruby, and searches for an introductory book. He/she does not know programming and therefore wants to search for a typical version of the book.

- A user plans to travel to Hokkaido and searches for a souvenir to buy there. He/she has already visited Hokkaido several times and bought some typical souvenirs; therefore, he/she now wants to find an atypical souvenir.

There are many other situations similar to the ones described above. Conventional search engines rank search result based on the relevance of each Web page to a query or based on the citation importance characterized by PageRank [7]. Therefore, it is difficult to search based on typicality. One way is to add the term "typical" to the original query. However, the keyword "typical" is not always included in a page containing typical information. Moreover, if typical information is written in a Web page, the user cannot judge whether the information is truly typical when he/she does not have knowledge about the domain.

In this research, we realize an information search on the basis of typicality. Although a great variety of information can be used as a search target, we target a set of objects (hereafter, "object set"). For example, a recipe and a tourist route are object sets because they are sets of ingredients and tourist spots, respectively. We propose a method for calculating the degree of typicality of an object set based on the appearance frequency of each object and the co-occurrence frequency between objects. However, there are various kinds of viewpoints that determine the degree of typicality of an object set. We follow the concept of typicality that is proposed in cognitive psychology, and propose methods for computing the degree of typicality of an object set for each viewpoint.

We also focus on search for an object set based on the addition or deletion of an object:

- A user plans to travel to Kyoto and is browsing a Web page that introduces a tourist route consisting of the Kiyomizu Temple, Nanzenji Temple, Heian-jingu Shrine, and Nijo-jo Castle. He is interested in the route but does not have enough time to visit all the attractions. Therefore, he wants to know which spots could be deleted from the route.

- A user plans to cook pasta carbonara and is browsing a Web page that introduces a pasta carbonara recipe. He is interested, but the recipe is too simple. Therefore, he wants to know what ingredient could be added to the recipe.

With respect to these search intentions, we recommend the addition or deletion object within an object set on the basis of the degree of typicality of the set. Specifically, we propose methods for recommending an addition and deletion object that results in a more typical or atypical object set. Although the target object set in this chapter is a recipe, our proposed method can be applied any kind of object set.

We conducted experiments using six categories: carbonara, napolitan, pork miso soup, minestrone, tomato salad, and tuna salad. In the experiments, we evaluated the correlation coefficient between the degree of typicality of recipes computed by the proposed method and that judged by assessors. We also evaluated the accuracy of the addition and deletion ingredients to a recipe recommended by the proposed method. Our results show the effectiveness of our method especially

in a category such as "minestrone" in which affinity between ingredients is important.

Additionally, we focus on two viewpoints of typicality, which are central tendency and frequency of instantiation, that were proposed in cognitive psychology. We target recipes and compare the typicality of a recipe judged by assessors with that calculated from each viewpoint.

3.2 Framework

Given a recipe o_u selected by a user, our system computes the typicality of the recipe, and recommends addition and deletion ingredients as follows:

(1) Detect the category c to which o_u belongs.

(2) Collect all recipes $O_c = \{o_1, o_2, ..., o_n\}$ in c.

(3) Collect all ingredients $E_c = \{e_1, e_2, ..., e_m\}$ each of which is used by at least one recipe in O_c.

(4) Detect the most typical set of ingredients, denoted by E_t, in c.

(5) Calculate the degree of typicality of o_u by comparing with E_t.

(6) Detect addition and deletion ingredients for o_u.

We focus on (4), (5), and (6), and propose methods in the following sections.

3.3 Methodology

3.3.1 The Most Typical Set of Objects in a Category

To calculate the typicality of a recipe, we first detect the most typical set of ingredients, denoted by E_t, in a category c. In this chapter, E_t is detected based on the appearance frequency of each ingredient and the co-occurrence frequency between ingredients. The co-occurrence frequency between ingredients e_i and e_j is defined by:

$$co(e_i, e_j) = \frac{|R(e_i) \cap R(e_j)|}{\min(|R(e_i)|, |R(e_j)|)}, \tag{3.1}$$

where $R(e_i)$ represents the set of recipes that include e_i in c.

E_t is detected as follows:

(1) Let S denote the ingredients in E_c whose $|R(e_i)|$ is higher than α.

(2) Set $E_t \leftarrow \phi$.

(3) Find the ingredient e_i that has the maximum value of $|R(e_i)|$ in S, and move it from S to E_t.

(4) Find the ingredient e_{max} in S that has the highest co-occurrence frequency with recipes that include all ingredients in E_t.

(5) If the co-occurrence frequency in step (4) is higher than β_1, set $S \leftarrow S \setminus \{e_{max}\}$ and $E_t \leftarrow E_t \cup \{e_{max}\}$, and go to step (4). Otherwise, regard E_t as the most typical set of ingredients.

3.3.2 Typicality of an Object Set

Given a recipe o and ingredients used in o, denoted by E_o, we calculate the typicality of a recipe o based on (1) the affinity between ingredients in E_o, and (2) the difference between E_t and E_o. The degree of typicality of E_o is calculated as follows:

$$f_{typ}(E_o) = f_{aff}(E_o) - f_{diff}(E_o, E_t), \ \ (-1 \leq f_{typ}(E_o) \leq 1) \tag{3.2}$$

where $f_{aff}(E_o)$ represents the affinity between ingredients in E_o, and $f_{diff}(E_o, E_t)$ represents the difference between E_t and E_o. The higher the value of $f_{typ}(E_o)$, the more typical the value of E_o. We describe the methods for calculating $f_{aff}(E_o)$ and $f_{diff}(E_o, E_t)$ in the remainder of this section.

$f_{aff}(E_o)$ is calculated based on the average co-occurrence frequency between ingredients in E_o as follows:

$$f_{aff}(E_o) = \frac{1}{{}_{|E_o|}C_2} \sum_{e_i, e_j \in E_o} co_1(e_i, e_j), \ \ (0 \leq f_{aff}(E_o) \leq 1) \tag{3.3}$$

where $co_1(e_i, e_j)$ is defined by:

$$co_1(e_i, e_j) = \begin{cases} 1 & co(e_i, e_j) > \theta \\ 0 & otherwise \end{cases} \tag{3.4}$$

Even if E_o includes only those ingredients that are used in multiple recipes in the category, $f_{aff}(E_o)$ has a low score when the ingredients are rarely combined with in the category. Conversely, even if E_o includes ingredients whose appearance frequency is low in the category, $f_{aff}(E_o)$ has a high score when those ingredients are often combined with in the category. In this paper, we set $\theta = 0.4$.

$f_{diff}(E_o, E_t)$ is calculated as follows:

$$f_{diff}(E_o, E_t) = (1-\mu) \sum_{e_i \in E_o \setminus T} \frac{1 - R'(e_i)}{|E_o \setminus T|} + \mu \frac{\sum_{e_i \in T \setminus E_o} R'(e_i)}{\sum_{e_i \in T} R'(e_i)}, \ \ (0 \leq f_{diff}(E_o, E_t) \leq 1) \tag{3.5}$$

where $R'(e)$ is defined by:

$$R'(e) = \frac{|R(e)|}{|R(e_{max})|}. \tag{3.6}$$

e_{max} is an ingredient that has the highest appearance frequency in the category. In Equation 3.5, the first member has a value between 0 and 1 when there are ingredients that are included in E_o but not in E_t. The value increases as more ingredients with low appearance frequencies are included in $E_o \setminus E_t$. That is, the more unusual ingredients included in E_o, the bigger the difference from E_t. The second member has a value between 0 and 1 when there are ingredients that are included in E_t but not in E_o. The value increases as more ingredients with high appearance frequencies are included in $E_o \setminus E_t$. That is, the fewer major ingredients included in E_o, the bigger the difference from E_t. In Equation 3.5, we set $\mu = 0.8$.

3.3.3 Candidate Addition and Deletion Objects

In this section, we describe the methods for detecting candidate addition and deletion ingredients for a recipe o in a category c.

Candidate Addition Objects

When we recommend addition ingredients to a recipe, the following two ingredient variables are recommended:

- An ingredient that changes the recipe to the most typical one by its addition.
- An ingredient that changes the recipe to the most atypical one by its addition.

However, we do not recommend an ingredient that changes the original recipe and makes it peculiar by adding it. Based on these conditions, our proposed method obtains addition ingredients as follows.

(1) Among E_c, we collect ingredients, denoted by E_f, whose value of $|R(e_i)|$ is higher than γ. Let $E_a = E_f \setminus E_o$ denote the candidate addition ingredients.

(2) For each ingredient in E_a, we calculate the degree of co-occurrence frequency with each ingredient in $E_o \cap E_f$ by using Equation 3.1. If the co-occurrence frequency of an ingredient is 0, it is removed from E_a.

(3) For each ingredient in E_a, we calculate the typicality of the recipe by adding it to E_o, and rank ingredients in E_a in descending order of score.

Table 3.1: Data for each category.

	carbonara	napolitan	pork miso soup	minestrone	tomato salad	tuna salad
Number of recipes	72	59	140	76	79	83
Number of all ingredients	581	576	1461	884	522	556
Number of unique ingredients	69	94	143	128	122	132
Average number of ingredients used in a recipe	8.07	9.76	10.4	11.6	6.61	6.70
Average number of same ingredients between two recipes	5.41	4.72	5.03	5.32	1.99	2.01
Average percentage of same ingredients between two recipes	67.1	48.3	48.2	45.7	30.1	30.1

Candidate Deletion Objects

When we recommend deletion ingredients in a recipe, the following two ingredients are recommended:

- An ingredient that changes the recipe to the most typical one by its deletion.
- An ingredient that changes the recipe to the most atypical one by its deletion.

Our proposed method obtains candidate deletion ingredients as follows.

(1) Following step (5) in Section 3.3.1, we collect a set of ingredients, denoted by E_b, in category c. Here, we use a threshold β_2 instead of β_1. Let $E_d = E_o \setminus E_b$ denote the candidate deletion ingredients.

(2) For each ingredient in E_d, we calculate the typicality of the recipe by deleting it from E_o, and rank ingredients in E_d in descending order of score.

3.4 Experiments

This section reports the evaluation of the proposed methods.

3.4.1 Dataset

We selected six categories in COOKPAD [1] for this experiment : "carbonara," "napolitan," "pork miso soup," "minestrone," "tomato salad," and "tuna salad." The number of recipes in each category was 72, 59, 140, 76, 79, and 83. We resolved the problem of inconsistency in ingredient spellings in advance by creating a dictionary. The ingredient data in each category is shown in Table 3.1.

3.4.2 Recipe Typicality

In this section, we evaluate the method proposed in Section 3.3.2 for calculating the typicality of a recipe.

[1] http://cookpad.com/

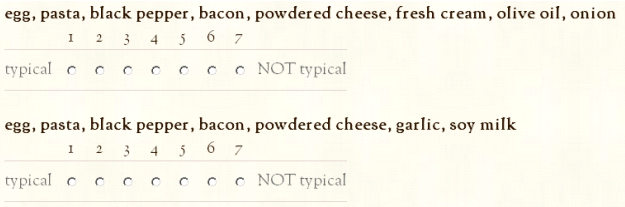

Figure 3.1: An example of questionnaire in the category of "carbonara."

Table 3.2: Rank correlation coefficient between the degree of typicality in the answer set and that calculated by the proposed method.

carbonara	napolitan	pork miso soup	minestrone	tomato salad	tuna salad
0.868	0.617	0.629	0.548	0.338	0.426

Answer Set

We regard the human-judged typicality of a recipe to be a correct answer. Three female assessors in their 20s generate an answer set. All of them were Kyoto University students who routinely cooked. To generate an answer set, we selected 40 recipes at random in each category. As shown in Figure 3.1, we showed the assessors only the ingredients used in each recipe. At the bottom of each set of ingredients, there was a 7-point Likert scale labeled from 1 (typical) to 7 (not typical), using which the assessors scored each set of ingredients. We asked them not to evaluate relatively between the 40 recipes, but between all existing recipes. Each recipe was evaluated by three assessors. We regard the average of their evaluations as the typicality of each recipe.

Results

Table 3.2 shows the rank correlation coefficient between the degree of typicality calculated by the proposed method and that in the answer set for each category.

In the categories of "carbonara," "napolitan," "pork miso soup," and "minestrone," the correlation coefficient was relatively high. As indicated in Table 3.1, the similarity of ingredients between recipes was relatively high in these categories. Here, the degree of recipe typicality computed by the proposed method was low when the recipe included ingredients with low appearance frequency or ingredients that were not usually combined. Assessors regarded such a recipe as atypical, and this resulted in high accuracies in these categories.

Conversely, in the categories of "tomato salad" and "tuna salad," the correlation coefficient was relatively low. As indicated in Table 3.1, the similarity of ingredients between recipes was moderate in these two categories, meaning that the appearance frequency of most ingredients

was low, and the number of ingredients that were often combined was also low. Hence, the proposed method, which calculates the degree of typicality based on the appearance frequency of ingredients and the co-occurrence frequency between ingredients, was not effective.

3.4.3 Addition and Deletion Ingredients

Procedure

In this experiment, we recruited eight male and two female assessors in their 20s and administered a questionnaire. Among them, three males and two females cooked routinely. We evaluated candidate addition and deletion ingredients recommended by the proposed method. We randomly selected nine recipes from each category. For each recipe, the following four ingredients were recommended by the proposed method:

(1) An ingredient that changes the recipe to a more typical one by its addition.

(2) An ingredient that changes the recipe to a more atypical one by its addition.

(3) An ingredient that changes the recipe to a more typical one by its deletion.

(4) An ingredient that changes the recipe to a more atypical one by its deletion.

For each recipe, the proposed method obtains ranked candidate addition ingredients as mentioned in Section 3.3.3. The top and bottom ranked ingredients were recommended in (1) and (2), respectively. Similarly, the proposed method obtains ranked candidate deletion ingredients. The top and bottom ranked ingredients were recommended in (3) and (4), respectively.

We use one baseline method that obtains candidate addition ingredients, as mentioned in Section 3.3.3. Here, the baseline method recommends the ingredient with the highest and lowest appearance frequencies in (1) and (2), respectively. Similarly, the baseline method obtains candidate deletion ingredients, as mentioned in Section 3.3.3, recommending the ingredient with the lowest and highest appearance frequency in (3) and (4), respectively.

For each recipe, ingredients used in the recipe were displayed to assessors. In addition, ingredients recommended by the proposed method and the baseline method in (1), (2), (3), and (4) were displayed.For each item, the assessors selected the most appropriate ingredient. When the same ingredient was recommended by the proposed and baseline method, the assessors could select both if they thought the ingredient was appropriate.

Results

Tables 3.3 and 3.4 list the evaluation results of assessors who cooked routinely and those who did not, respectively.

Table 3.3: Percentage of ingredients selected by subjects who cook routinely in each category.

		addition typical	addition atypical	deletion typical	deletion atypical
carbonara	proposed method	84.4%	64.4%	71.1%	26.7%
	baseline method	84.4%	37.8%	62.2%	26.7%
	not selected	15.6%	4.4%	20.0%	73.3%
napolitan	proposed method	82.2%	68.9%	55.6%	46.7%
	baseline method	80.0%	31.1%	55.6%	57.8%
	not selected	13.3%	0.0%	8.9%	31.1%
pork miso soup	proposed method	68.9%	44.4%	71.1%	48.9%
	baseline method	88.9%	55.6%	57.8%	51.1%
	not selected	4.4%	0.0%	11.1%	44.4%
minestrone	proposed method	84.4%	71.1%	75.6%	33.3%
	baseline method	64.4%	28.9%	44.4%	8.9%
	not selected	8.9%	0.0%	4.4%	62.2%
tomato salad	proposed method	68.9%	53.3%	46.7%	33.3%
	baseline method	48.9%	46.7%	64.4%	35.6%
	not selected	2.2%	0.0%	13.3%	57.8%
tuna salad	proposed method	71.1%	51.1%	66.7%	35.6%
	baseline method	42.2%	46.7%	71.1%	35.6%
	not selected	11.1%	2.2%	24.4%	53.3%
average	proposed method	76.7%	58.9%	64.4%	37.4%
	baseline method	68.1%	41.1%	59.3%	35.9%
	not selected	9.3%	1.1%	13.7%	53.7%

In the results of the assessors who cooked routinely, the average ratio of the proposed method outperformed the baseline method in all items. The proposed method was especially effective in changing an original recipe to a more atypical one by adding an ingredient. In the category of "pork miso soup," however, the baseline method outperformed the proposed method in "addition atypical." For pork miso soup, the affinity between ingredients is not a problem because there are only a few styles, such as Chinese or Western style. The assessors assumed that a recipe becomes atypical simply by adding a rare ingredient, and therefore the baseline method that considered only the appearance frequency of ingredients outperformed the proposed method. Conversely, in the categories of "carbonara," "napolitan," and "minestrone," there are various kinds of style, according to ingredients used such as vegetables and flavoring materials. Hence, the proposed method that considers the affinity between ingredients worked better. In the categories of "tomato salad" and "tuna salad," too, the affinity between ingredients is important; there are too many recipe styles. Hence, the degree of typicality of a recipe varied from one assessor to another, and there were few differences between the two methods.

There were no major differences in "addition typical" and "deletion typical" between Ta-

Table 3.4: Percentage of ingredients selected by subjects who do not cook routinely in each category.

		addition typical	addition atypical	deletion typical	deletion atypical
carbonara	proposed method	75.6%	51.1%	73.3%	28.9%
	baseline method	75.6%	42.2%	60.0%	28.9%
	not selected	24.4%	8.9%	22.2%	71.1%
napolitan	proposed method	88.9%	35.6%	57.8%	62.2%
	baseline method	82.2%	64.4%	48.9%	68.9%
	not selected	8.9%	0.0%	24.4%	17.8%
pork miso soup	proposed method	68.9%	53.3%	66.7%	66.7%
	baseline method	82.2%	46.7%	57.8%	80.0%
	not selected	8.9%	0.0%	15.6%	20.0%
minestrone	proposed method	75.6%	37.8%	68.9%	37.8%
	baseline method	73.3%	60.0%	44.4%	17.8%
	not selected	13.3%	2.2%	8.9%	53.3%
tomato salad	proposed method	66.7%	48.9%	68.9%	28.9%
	baseline method	55.6%	51.1%	68.9%	35.6%
	not selected	0.0%	0.0%	4.4%	64.4%
tuna salad	proposed method	68.9%	51.1%	84.4%	35.6%
	baseline method	46.7%	48.9%	77.8%	40.0%
	not selected	8.9%	0.0%	13.3%	51.1%
average	proposed method	74.1%	46.3%	70.0%	43.3%
	baseline method	69.3%	52.2%	59.6%	45.2%
	not selected	10.7%	1.9%	14.8%	46.3%

ble 3.3 and 3.4. This indicates that even assessors who do not cook routinely were able to select an appropriate ingredient to change an original recipe to a more typical one. However, in Table 3.4, the baseline method outperformed the proposed method in "addition atypical" and "deletion atypical" in many categories. This means that assessors who do not cook routinely selected an ingredient mainly based on the rarity. Therefore, our proposed method is especially useful for users who do not cook routinely when they want to change a recipe to a more atypical one by ingredient addition or deletion.

3.5 Analysis of Typicality

In this section, we analyze typicality of object sets from *central tendency* (CT) and *frequency of instantiation* (FOI) concepts. First, we describe methods for computing typicality of an object set from each concept. We then conduct experiments to analyze the relation between human-judged typicality and typicality from each concept. We also discuss the relation between our proposed method in Section 3.3 and each concept.

3.5.1 Typicality from Two Viewpoints

Typicality based on Central Tendency

In central tendency, an object is typical when it is similar to many other objects in the category. In this paper, we regard the similarity of objects as the similarity of property, and use the TextRank algorithm [44] to calculate it. In the TextRank algorithm, a sentence is represented by a vector in which each element is a term. This algorithm can detect the most important sentence on the basis of the similarity between sentences. We can apply the TextRank algorithm to a set of objects because each object can be represented by a vector.

We follow Yeung and Leung [79] and consider an object as a $property\ vector$. The property vector of an object o in a category c is represented by a vector of property:value pairs.

$$\boldsymbol{p}_o = (p_{o,1} : v_{o,1}, p_{o,2} : v_{o,2}, \cdots, p_{o,k} : v_{o,k}), 0 \leq v_{o,i} \leq 1, \tag{3.7}$$

where k is the total number of properties in the category, and $v_{o,i}$ indicates the fuzzy degree to which the object o in category c possesses the property $p_{o,i}$. In the category of "bird," for example, a bird o is represented as follows:

$$\boldsymbol{p}_o = (Animal : 1, Has` - Wings : 1, \cdots, Can` - Run : 0.2) \tag{3.8}$$

When we apply the TextRank algorithm to a set of recipes, a graph is made in which a recipe is a node and the similarity between two recipes is the weight of an edge. Hence, we can regard the score of each recipe as the similarity between whole recipes. The TextRank algorithm is calculated by a recursive calculation as follows:

$$\boldsymbol{TR} = \alpha \cdot \boldsymbol{S}^* \times \boldsymbol{TR} + (1 - \alpha) \cdot \boldsymbol{p},\ where\ \ \boldsymbol{p} = [\frac{1}{n}]_{n \times 1}, \tag{3.9}$$

where $\boldsymbol{S}^*$ is a normalized matrix of a similarity function matrix $\boldsymbol{S}$ that represents the similarity between recipes, and $\boldsymbol{TR}$ is the typicality of recipes. Then $\boldsymbol{p}$ is a vector representing the probability of choosing a recipe randomly without following an edge between recipes, and α is a $dampingfactor$. The ranking of each recipe score after applying TextRank to the recipes is the ranking of typicality from this viewpoint.

Typicality based on Frequency of Instanciation

In frequency of instantiation, one factor used to determine the typicality of an object is that an object with a higher cognition is more typical. There are ways to estimate the typicality from information on the Web. One is the ranking of each Web page in which each object is

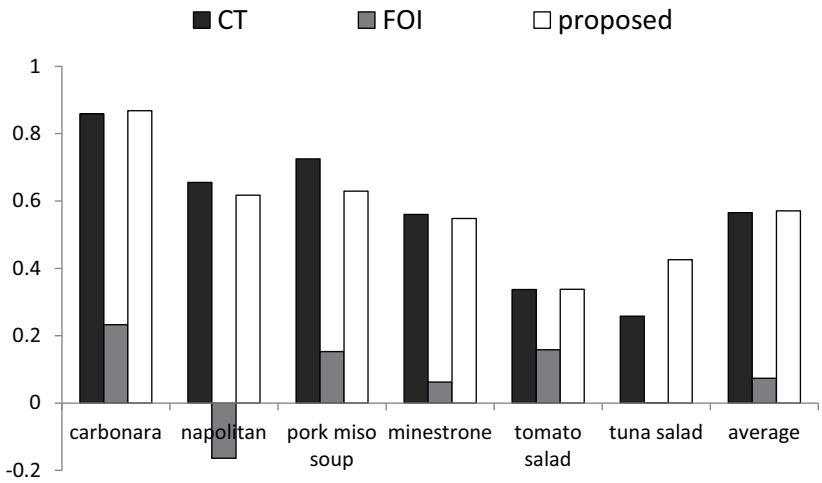

Figure 3.2: Rank correlation coefficient between the typicality that the assessor labeled and that based on each viewpoint and our proposed method (y-axis: rank correlation coefficient).

included in a search result. High-ranked Web pages are read by many general users; consequently, objects in these Web pages have a high degree of cognition. For example, in the category, "Kyoto sightseeing spot," people think "Kinkakuji" is typical. This is because a Web page including "Kinkakuji" tends to be ranked high when a user searches using the query, "Kyoto sightseeing spot." Another way is by the number of Web pages including each object in a search result. That is, "Kinkakuji" is included in more Web pages than other sightseeing spots when a user searches with the query, "Kyoto sightseeing spot." The number of social bookmarks for each Web page is also a criterion for cognition degree. In the case of recipes uploaded to COOKPAD, not all recipes are included in search results when we search the Web with the category name. In COOKPAD, there is a system called "tsukurepo." This is a system in which a user who has used a recipe uploaded by other users posts a recipe report. A recipe with many tsukurepo reports is ranked high when a user searches for recipes; therefore, such a recipe has high recognition. In this paper, we regard a recipe with more tsukurepo reports as more typical, and regard the ranking of the number of tsukurepo reports as the ranking of degree of typicality.

3.5.2 Experiments

We performed an experiment to survey the relation between human-judged typicality and that based on each of the two viewpoints. We used the same answer set as in Section 3.4.2 and computed the rank correlation coefficient between the typicality calculated in each viewpoint and that in the answer set in each category.

The results are shown in Figure 3.2. This figure also shows the rank correlation coefficients

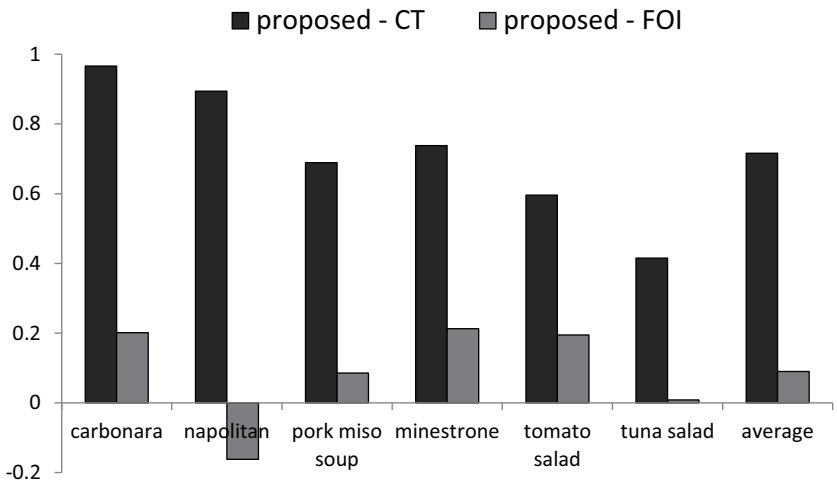

Figure 3.3: Rank correlation coefficient between the typicality based on each viewpoint and that computed by our proposed method (y-axis: rank correlation coefficient).

between human-judged typicality and that computed by our proposed method in Section 3.3. As shown in Figure 3.2, our original proposed method marked the highest rank correlation coefficient on average. Central tendency also scored a high rank correlation coefficient; however, the minimal value was low compared to our proposed method. In the category of "carbonara," in which the similarity of ingredients between recipes was high, typicality based on central tendency was especially effective. The minimal value was scored in the category of "tuna salad," a category in which there were few commonly used ingredients; therefore, central tendency was not efficient. However, our method showed robustness even in such a case. Frequency of instantiation marked a low rank correlation coefficient in all categories. This means that popular recipes are not always typical recipes; however, by using the viewpoint of frequency of instantiation, we can find a recipe that is not typical in the viewpoint of central tendency, but is popular.

Finally, Figure 3.3 shows rank correlation coefficients between the typicality based on each viewpoint and that computed by our proposed method. In the categories of "carbonara," "napolitan," "pork miso soup," and "minestrone," rank correlation coefficients were especially high between our proposed method and central tendency, while rank correlation coefficients were especially high between our proposed method and frequency of instantiation were low in all categories. That is, we can deduce that our proposed method is a central tendency oriented method.

Although we only considered ingredients for computing the degree of typicality of recipes, when users choose a recipe, they usually not only consider the ingredients but also step-by-step instructions, images, recipe creators, and so on. Therefore, to compare the typicality of recipes more accurately, we should consider such factors and propose suitable methods.

3.6 Summary

In this chapter, we focused on an object set such as a recipe or a tourist route, which consists of some objects, such as ingredients and tourist spots, and proposed a method for calculating the degree of typicality of the object set. The proposed method first detects the most typical set of objects in a category based on the appearance frequency of each object and the co-occurrence frequencies between objects. Given an object set, we compute the degree of its typicality based on the affinity between its objects and the difference between the object set and the most typical set of objects. We also proposed methods for recommending candidate addition and deletion objects to an object set to change it to a more typical or atypical one.

We focused on recipes as object sets and conducted experiments. The results showed that the correlation coefficient between the typicality judged by assessors and that computed by the proposed method was as high as 0.868 in a category. In the experiment regarding addition and deletion of ingredients, we found that the proposed method was especially effective in recommending addition and deletion ingredients to change a recipe to a more atypical one.

We also focused on each viewpoint of typicality, as proposed in cognitive psychology. We targeted recipes and proposed methods for calculating typicality for each viewpoint. Evaluation experiments showed that a viewpoint based on similarity was able to estimate the typicality judged by assessors with high accuracy in a category in which the similarity of properties between objects was high.

We plan to evaluate the general versatility of the proposed method by applying it to object sets other than recipes.

CHAPTER 4

RANKING OF COORDINATE TERMS AND HYPERNYMS USING A HYPERNYM-HYPONYM DICTIONARY

4.1 Introduction

Given a term t, there are various types of relationships between t and other terms. For example, hypernyms and hyponyms are defined as terms that are more general and specific than t, respectively. A synonym is a term with the same meaning for t as another term, and a coordinate term is a term that has one or more common hypernyms with t. There are also other relationships such as antonyms and related terms. This study focuses on hypernyms and coordinate terms to identify appropriate hypernyms and coordinate terms for a given query.

Discovering coordinate terms for a given query is useful in various situations. For instance, suppose a user inputs a query to a Web search engine, and is not familiar with Web search or does not have sufficient knowledge about the search domain. In such a case, displaying coordinate terms of the query would support his Web search. For example, if a user needs information about digital cameras but knows only "LUMIX," then displaying appropriate coordinate terms, such as "EXLIM," "FinePix," and "Cyber-Shot" for comparison may be useful to him. Similarly, discovering hypernyms of terms is also useful in some situations such as connecting diverse concepts to form a semantic taxonomy [69].

Some studies have proposed methods for discovering coordinate terms or hypernyms of a term [29, 35, 55, 62, 69, 70, 76, 78]. The aim of these studies is only discovering these terms from unstructured data such as Web pages and query logs of a commercial search engine. The studies evaluate discovered hypernyms or coordinate terms based on a binary evaluation: appro-

priate or inappropriate. In this research, we use a hypernym-hyponym dictionary (described in Section 4.2.1), which enables us to easily obtain hypernyms and coordinate terms of a given term. However, from the dictionary we obtain a large number of coordinate terms and hypernyms. For example, for the query "Lionel Messi," we obtain 16 hypernyms and 112,489 coordinate terms from the dictionary. However, as will be described in Section 4.2.2, there are appropriate and inappropriate hypernyms as well as coordinate terms among these results. Thus, although both "Cristiano Ronaldo" and "Stevie Wonder" are coordinate terms of "Lionel Messi," "Cristiano Ronaldo" is more appropriate than "Stevie Wonder." Similarly, for "Lionel Messi," "football player" is a more appropriate hypernym than "human beings."

In this research, we propose methods for ranking coordinate terms and hypernyms of a term based on their appropriateness. Our method first creates a bipartite graph based on hypernyms of a query and hyponyms of each hypernym using a hypernym-hyponym dictionary. We apply a HITS-based algorithm to the graph and rank coordinate terms and hypernyms based on their appropriateness. Although we use a Japanese hypernym-hyponym dictionary, our methods are language-independent.

The experimental results obtained using 50 queries demonstrate that our method can rank appropriate coordinate terms and hypernyms higher than other comparable methods.

The contributions of this study are twofold:

- We propose methods for ranking coordinate terms and hypernyms by considering their appropriateness. Most of the previous studies have focused on only discovering coordinate terms and hypernyms for a given query, whereas our objective is the ranking of coordinate terms and hypernyms according to their appropriateness.

- We evaluate coordinate terms and hypernyms based on their appropriateness. Most previous studies have evaluated discovered coordinate terms and hypernyms based on a binary evaluation, whereas we evaluate coordinate terms and hypernyms by considering their appropriateness.

The remainder of the chapter is organized as follows. In Section 4.2, we describe the hypernym-hyponym dictionary used in this study and our proposed method. In Section 4.3, we report our evaluation experiments. In Section 4.4, we discuss the results obtained. Finally, in Section 4.5, we provide our conclusion and present possible suggestions for future studies.

4.2 Method

In this section, we describe the hypernym-hyponym dictionary used in this study, discuss the characteristics of appropriate coordinate terms and hypernyms, and present methods to rank these

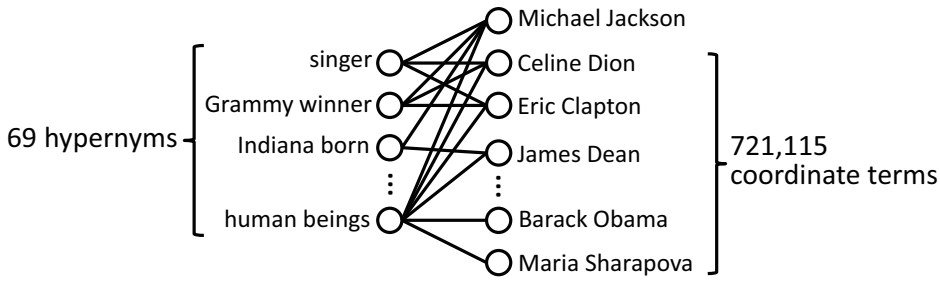

Figure 4.1: Examples of Michael Jackson's hypernyms and coordinate terms.

coordinate terms and hypernyms.

4.2.1 Hypernym-Hyponym Dictionary

In this research, we use an open source "hypernym/hyponym extraction tool [1]." This tool contains approximately 200,000 hypernyms and approximately 2.45 million hyponyms. These hierarchized terms are category names and nouns that occur in the titles of articles in Japanese Wikipedia. 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 69 hypernyms such as "singer" and "Guinness world record holder." Thus, if a term has at least one common hypernym with "Michael Jackson," then the term is a coordinate term of "Michael Jackson," and "Michael Jackson" has 721,115 coordinate terms (Figure 4.1).

4.2.2 Characteristics of Appropriate Coordinate Terms and Appropriate Hypernyms

In this research, we define a coordinate term of a term q as "a term that has one or more common hypernyms with q," as defined by Ohshima *et al.* [55]. Similarly, a hypernym of a term is defined as "a term that is more general than q." However, among coordinate terms and hypernyms obtained using the aforementioned dictionary, there are gaps in the degrees of appropriateness of coordinate terms and hypernyms.

First, we studied the characteristics of appropriate coordinate terms of q and found the following characteristics:

(1-A) An appropriate coordinate term shares many hypernyms with q.

(1-B) An appropriate coordinate term shares hypernyms that have fewer hyponyms with q.

We explain these characteristics using "Lionel Messi" as an example. Thus, given two terms, "Cristiano Ronaldo" and "Stevie Wonder," "Cristiano Ronaldo" is a more appropriate coordinate term of "Lionel Messi," which can be explained by considering (1-A). In this case, "Stevie

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

Wonder" shares only one hypernym, "human beings," with "Lionel Messi," whereas "Cristiano Ronaldo" shares both "human beings" and "football player." Similarly, given two additional terms, "Wayne Rooney" and "Jorge Luis Borges," "Wayne Rooney" is more appropriate as a coordinate term of "Lionel Messi." However, when we consider only (1-A), the appropriateness of these two terms is equivalent, because "Wayne Rooney" shares two hypernyms, "human beings" and "football player," with "Lionel Messi" and "Jorge Luis Borges" shares two hypernyms, "human beings" and "from Argentina." Hence, in this case, the difference can be explained by considering (1-B); the number of "football player" is fewer than "from Argentina." Therefore "Wayne Rooney" is more appropriate as a coordinate term.

Second, we studied the characteristics of appropriate hypernyms of q and determined the following characteristics:

(2-A) An appropriate hypernym has only appropriate coordinate terms of q as its hyponyms.

(2-B) An appropriate hypernym has many appropriate coordinate terms of q as its hyponyms.

We will also explain these characteristics using "Lionel Messi" as an example. Thus, given two hypernyms, "football player" and "human beings," "football player" is a more appropriate hypernym of "Lionel Messi," which can be explained by considering (2-A). In this case, "human beings" has appropriate coordinate terms of "Lionel Messi" such as "Cristiano Ronaldo" and the inappropriate coordinate terms such as "Stevie Wonder" and "Barack Obama," but "football player" only has appropriate coordinate terms such as "Cristiano Ronaldo" and "Wayne Rooney." Similarly, given two additional terms, "football player" and "winner of UEFA Best Player in Europe Award," we think "football player" is more appropriate as a hypernym of "Lionel Messi" because "winner of UEFA Best Player in Europe Award" is too narrow as a hypernym. However, when we consider only (2-A), the appropriateness of these two terms is equivalent, because both hypernyms have only appropriate coordinate terms of "Lionel Messi" as their hyponyms. Hence, in this case, the difference can be explained by considering (2-B); "football player" has more coordinate terms of "Lionel Messi" as its hyponyms than "winner of UEFA Best Player in Europe Award." Therefore, "football player" is more appropriate as a hypernym of "Lionel Messi."

4.2.3 Ranking of Coordinate Terms

First we will define some symbols. Let q denote a query and $hyper(t)$ and $hypo(t)$ denote the set of hypernyms and set of hyponyms of a term t, respectively. H_q and C_q are defined as follows.

- $H_q = \{x | x \in hyper(q)\}$,
- $C_q = \{x | x \in hypo(y), y \in H_q, x \neq q\}$.

That is, H_q and C_q are the set of hypernyms and the set of coordinate terms of q, respectively.

We consider a bipartite graph $G = (\{q\} \cup C_q \cup H_q, E)$, where E is a set of edges between H_q and $\{q\} \cup C_q$. An edge exists between $h_i \in H_q$ and $c_j \in \{q\} \cup C_q$ when h_i is a hypernym of c_j. When q is "Michael Jackson," Figure 4.1 represents the bipartite graph.

To calculate the appropriateness of each coordinate term in C_q, we propose a method that reflects characteristics (1-A) and (1-B) based on the HITS [36] algorithm. Originally the HITS algorithm was used to evaluate Web pages based on link structure. 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 $h_i \in H_q$ and the authority score of $c_j \in \{q\} \cup C_q$ as $hub(h_i)$ and $authority(c_j)$, respectively, and calculate these scores as follows:

$$hub(h_i) = \sum_{c_j \in \{q\} \cup C_q} w_{ji}^{ch} \cdot authority(c_j), \tag{4.1}$$

$$authority(c_j) = \sum_{h_i \in H_q} w_{ij}^{hc} \cdot hub(h_i), \tag{4.2}$$

where w_{ji}^{ch} and w_{ij}^{hc} represent the weight of edges, and w_{ji}^{ch} represents the weight from c_j to h_i. In the HITS algorithm, the weight of an edge is equal to 1 if there is an edge between two vertices, otherwise the weight of an edge is equal to 0. If we apply the HITS algorithm to the bipartite graph G then vertices that have a very large number of hyponyms, such as "human beings" and "from Argentina," have high scores. Thus, each hyponym of "human beings" or "from Argentina" has a misleading high score, and terms sharing hypernyms that have many hyponyms become appropriate coordinate terms of q. To solve this problem, we change the weight of edges from hypernyms to hyponyms by considering the number of hyponyms of each hypernym as mentioned in (1-B). Lempel and Moran [37] proposed the SALSA algorithm, considering the weight of edges in the HITS algorithm. In the SALSA 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 c_j is represented by $w_{ij}^{hc} = \frac{1}{|hypo(h_i)|}$.

We set the initial value of q as 1 and the initial values of the remaining vertices as 0, because the objective of our method is to calculate the degree of coordination to q. Let $f_{coordinate}(q, c_j)$ and $f_{multitude}(q, h_i)$ denote the convergent scores of $c_j \in C_q$ and $h_i \in H_q$, respectively. When we rank coordinate terms of q based on their appropriateness, we sort $c_j \in C_q$ in descending order of $f_{coordinate}(q, c_j)$.

Table 4.1: Examples of queries (English translation).

category	queries
person	Paul McCartney, Tom Cruise, Ichiro Suzuki, Ludwig van Beethoven, Nobunaga Oda
place	United Kingdom, Paris, Tokyo, the Pacific Ocean, Brazil
product	digital camera, Nintendo DS, refrigerator, frying pan, organ
facility	department store, the University of Tokyo, Universal Studios Japan, Narita International Airport
company	Microsoft Corporation, Panasonic, McDonald's Corporation, Adidas, Toyota Motor Corporation

4.2.4 Ranking of Hypernyms

The score $f_{multitude}(q, h_i)$ reflects only characteristic (2-B) from Section 4.2.2. Therefore, hypernyms such as "human beings" have high scores.

To reflect characteristic (2-A), "an appropriate hypernym has *only* appropriate coordinate terms of q as its hyponyms," we calculate the score of $h_i \in H_q$ as follows:

$$f_{purity}(q, h_i) = \frac{1}{|hypo(h_i)|} \sum_{t_j \in hypo(h_i)} f_{coordinate}(q, t_j) \ . \tag{4.3}$$

That is, $f_{purity}(q, h_i)$ is the average score of the degree of coordination for all of h_i's hyponyms. Finally the appropriateness score of h_i as a hypernym of q is given by:

$$f_{hypernym}(q, h_i) = f_{purity}(q, h_i)^{\beta} \cdot f_{multitude}(q, h_i)^{(1-\beta)} \ , \tag{4.4}$$

where β is a parameter that ranges from 0 to 1.

4.3 Experiments

This section reports on the evaluation of the proposed methods.

4.3.1 Query Set

We created a query set comprising 50 queries in five categories: names of people, places, products, facilities, and companies. Each category contains ten queries. These queries are Wikipedia pages, where the title of the page is the query. If a query is unpopular, evaluating is difficult for assessors. Therefore, we have selected popular queries as follows. First, we compute PageRank [7] scores for all Wikipedia articles based on their link structures. Queries with high PageRank scores are considered popular, and we then select the top 100 queries for each category. Finally, we randomly select ten popular queries for each category. Examples from the query set are presented in Table 4.1.

4.3.2 Comparative Methods

Coordinate term

In this experiment, two comparative methods were used to compute the degree of coordination. The first method, denoted the CommonHypernym method, hypothesizes that the more hypernyms a term $c_j \in C_q$ shares with a query q, the higher the degree of coordination of c_j becomes. That is, the score of $c_j \in C_q$ is calculated as follows:

$$f_{common\_hypernym}(q, c_j) = |hyper(q) \cap hyper(c_j)| \tag{4.5}$$

The second method, denoted the SALSA method, sets $w_{ji}^{ch} = \frac{1}{|hyper(c_j)|}$ and $w_{ij}^{hc} = \frac{1}{|hypo(h_i)|}$ in Equation 4.2. The SALSA method hypothesizes that the fewer hypernyms a term $c_j \in C_q$ shares with q and the fewer hyponyms each of the hypernyms have, the more appropriate coordinate term c_j is. More intuitively, a term that shares only rare hypernyms with q is an appropriate coordinate term of q.

Hypernym

Two comparative methods were used to compute the hypernym score. The first method, denoted the ManyHyponyms method, hypothesizes that the more hyponyms a hypernym $h_i \in H_q$ has, the more appropriate hypernym h_i is: i.e., the appropriateness score of h_i is calculated by $|hypo(h_i)|$.

In contrast, the second method, denoted the FewHyponyms method, hypothesizes that the fewer hyponyms a hypernym $h_i \in H_q$ has, the more appropriate hypernym h_i is: i.e., the appropriateness score of h_i is calculated by $\frac{1}{|hypo(h_i)|}$.

4.3.3 Evaluation Procedure

Evaluation of Coordinate Terms

For a given a query, the proposed method and two comparative methods can calculate the degree of coordination for all coordinate terms of the query. However, the average number of coordinate terms for queries used in this experiment was extremely high (263,143.98 terms per query). Manually evaluating the degree of coordination of all terms is difficult; thus, for a given query, we pooled the top 50 coordinate terms from each method to solve this problem. The pooled terms were then randomly sorted and evaluated.

Assessors were recruited through Lancers [2], which is a popular crowd sourcing marketplace in Japan. First, we presented a query and asked the assessors to label each of the query's coordinate terms from 0 to 2, where 0 indicates that the term is not appropriate as the coordinate term, 1 indicates that the term is reasonably appropriate, and 2 indicates that the term is absolutely

[2] http://www.lancers.jp/

Table 4.2: MAP of coordinate terms. The highest scores in each category are indicated in bold. Paired t-tests with Bonferroni corrections were used for significance testing. Significant differences between the proposed method and CommonHypernym are indicated by $*(\alpha = 0.05)$, and significant difference between the proposed method and SALSA is indicated by † † $(\alpha = 0.01)$.

	CommonHypernym	SALSA	Proposed
person	0.535	0.557	**0.578**
place	0.505	0.535	**0.549**
product	0.425	0.468	**0.548***
facility	0.701	0.646	**0.714**
company	0.637	0.601	**0.651**
all categories	0.561	0.560	**0.608***††

appropriate. If the assessors were not able to attribute a score for a coordinate term because they did not understand the term, we asked them to label it "unknown" rather than attributing a score. Each coordinate term was labeled by 11 assessors.

Evaluation of Hypernyms

For hypernyms, the average number of hypernyms of queries used in this experiment was reasonable (46.4 hypernyms per query). Thus, we used all hypernyms of the queries in this experiment. Again, we used Lancers to recruit assessors. Initially, we displayed a query and asked the assessors to label each of its hypernyms on a scale from 0 to 2. For a given hypernym, 0 indicates that the term is not appropriate, 1 indicates that the term is reasonably appropriate, and 2 indicates that the term is absolutely appropriate. If the assessors were not able to label the score for a hypernym because they did not understand the term, we asked them to label it "unknown" rather than attributing a score. Each hypernym was labeled by 11 assessors.

4.3.4 Evaluation Metrics

We used Normalized Discounted Cumulated Gain (nDCG) [32] and Mean Average Precision (MAP) as evaluation metrics. To compute both metrics for coordinate terms, we first listed coordinate terms that more than seven assessors had labeled "unknown." Hereafter, we denote such terms "unknown terms." As mentioned previously, each of the three methods has a term list of the top 50 ranked coordinate terms. Unknown terms were discarded from the list, and the remaining coordinate terms were re-ranked according to their degrees of coordination. Then, we computed the average assessor scores for each remaining coordinate term and regarded this score as the answer score. To compute both metrics for hypernyms, we followed a similar procedure and computed the answer score for each hypernym.

To compute the MAP for coordinate terms, the coordinate terms must be divided into two

Table 4.3: Comparison of nDCG among all methods. The highest scores at each rank are shown in bold. Paired t-tests with Bonferroni corrections were used for significance testing. Significant differences between the proposed method and CommonHypernym are indicated by $*(\alpha = 0.05)$.

	CommonHypernym	SALSA	Proposed
@5	0.709	0.709	**0.743**
@10	0.713	0.715	**0.747***
@20	0.732	0.739	**0.762***
@30	0.769	0.774	**0.793**

Table 4.4: nDCG for each category computed by the proposed method.

	person	place	product	facility	company
@5	0.734	0.705	0.689	0.773	0.814
@10	0.766	0.709	0.694	0.766	0.799
@20	0.798	0.745	0.700	0.779	0.787
@30	0.823	0.789	0.731	0.818	0.803

groups: appropriate and inappropriate coordinate terms. In this experiment, we considered coordinate terms with answer scores ≥ 1 as appropriate, while coordinate terms with answer scores < 1 were treated as inappropriate terms. Hypernyms were treated in the same manner and were also divided into two groups: appropriate and inappropriate.

4.3.5 Results

Results of Coordinate Terms

Table 4.2 presents the MAP for each category. Paired t-tests with Bonferroni corrections were used for significance testing. The proposed method significantly outperformed both the CommonHypernym and SALSA methods for the average of 50 queries. Moreover, the proposed method outperformed other comparable methods in all five categories.

Table 4.3 presents a comparison of nDCG for all methods. Although nDCG at rank 40 or 50 cannot be calculated for some queries because unknown terms were discarded (See Section 4.3.4), the nDCG at rank ≤ 30 can be calculated for all queries. Thus, the nDCG at rank 5, 10, 20 and 30 are presented in Table 4.3. Paired t-tests with Bonferroni corrections were used for significance testing. The results obtained indicate that the proposed method achieved the highest nDCG at any rank (from nDCG@5 to nDCG@30), and this method significantly outperformed the CommonHypernym method at rank 10 and 20.

Table 4.4 presents the nDCG for queries from each category computed by the proposed

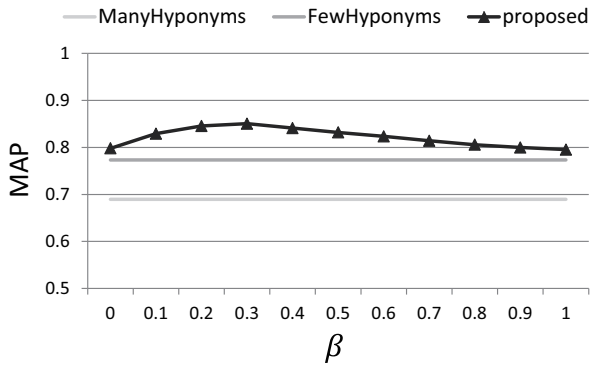

Figure 4.2: MAP for the average of 50 queries in each method (β ranges from 0 to 1 in increments of 0.1).

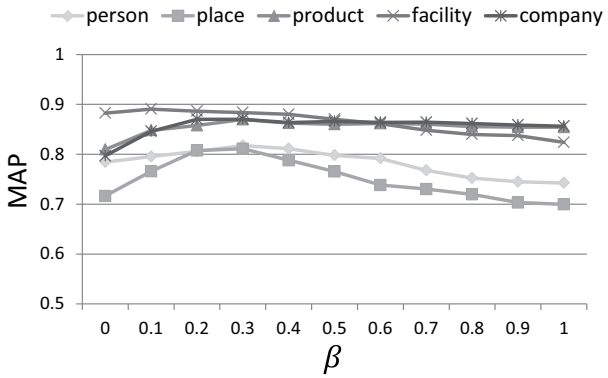

Figure 4.3: MAP in each category (β ranges from 0 to 1 in increments of 0.1).

method. From the results in Table 4.4, we can say that the proposed method was effective in the person, facility, and company categories, and less effective in product category.

Results of Hypernyms

Figure 4.2 presents MAP result comparisons for the average of 50 queries for all methods. The proposed method has a parameter β, which ranges from 0 to 1 in increments of 0.1. Two comparative methods have no parameter, and have scored constant MAP values regardless of β. Figure 4.2 determines that the proposed method outperformed two comparative methods for any value of β. The proposed method achieved the highest value (0.850) when β was 0.3, indicating the effectiveness of considering the characteristics of both (2-A) and (2-B) from Section 4.2.2.

Figure 4.3 illustrates the MAP for each category when β ranged from 0 to 1 in increments of 0.1. The MAP achieved the highest value when β was 0.1 in the facility category and 0.3 in other categories.

Figure 4.4 presents the average nDCG for the average of 50 queries when β ranged from 0

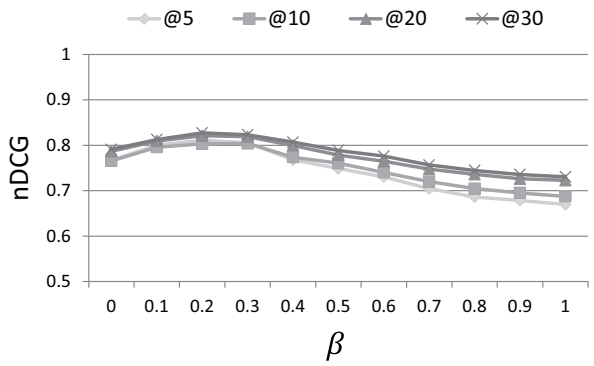

Figure 4.4: nDCG of all queries. (β ranges from 0 to 1 in increments of 0.1).

to 1 in increments of 0.1. At any rank, the nDCG achieved the highest value when β was 0.3 (@5, @20, and @30) or 0.4 (@10). These results indicates the effectiveness of combining the characteristics of both (2-A) and (2-B) from Section 4.2.2.

4.4 Discussion

In this section, we discuss the results with some specific examples.

4.4.1 Coordinate Terms

Table 4.5 presents the results for an example in which the proposed method determined appropriate coordinate terms with high accuracy. Each column of the table displays the top 20 coordinate terms of the CommonHypernym method, SALSA method, and the proposed method, as well as the top 20 terms in terms of answer scores.

Table 4.5 shows the results for the query "Paul McCartney." In the answer data, famous western singers were regarded as appropriate coordinate terms of "Paul McCartney," and the proposed method ranked such terms higher. In the CommonHypernym method, terms that do share many hypernyms with the query were ranked higher. However, the method does not consider the importance of each hypernym. Terms such as "Keisuke Kuwata" and "Tsuyoshi Nagabuchi," which are names of famous Japanese singers, were labeled as inappropriate coordinate terms by many assessors and were ranked higher in the CommonHypernym method. They share unimportant hypernyms such as "a singer who plays different instruments when playing different forms of music" with "Paul McCartney." The SALSA method also placed high priority on such hypernyms; thus, the nDCG was lower than that of the proposed method.

According to our observations, there are two principal cases when our methods did not work efficiently. The first case is for a query that has multiple meanings. For example, "Japan Sea" has totally 31 hypernyms. Among the hypernyms of the query, 13 hypernyms are related to a

Table 4.5: Ranking results for coordinate terms from the proposed method and comparison methods for the query "Paul McCartney" (numbers in the parentheses indicate answer score).

rank	CommonHypernym	SALSA	Proposed	answer data
1	Elton John	Elton John	Eric Clapton	John Lennon (1.82)
2	Eric Clapton	Sting	Ringo Starr	BEATLES (1.82)
3	Sting	Eric Clapton	Elton John	Ringo Starr (1.70)
4	John Lennon	John Lennon	John Lennon	Michael Jackson (1.50)
5	Keisuke Kuwata	Ringo Starr	David Bowie	George Harrison (1.45)
6	Mariah Carey	Keisuke Kuwata	BEATLES	Linda McCartney (1.43)
7	Stevie Wonder	Mariah Carey	Sting	Elton John (1.29)
8	Mick Jagger	Stevie Wonder	Celine Dion	Wings (1.25)
9	Paul Simon	George Harrison	Mariah Carey	Stevie Wonder (1.20)
10	Tsuyoshi Nagabuchi	Mick Jagger	George Harrison	Prince (1.14)
11	Keith Richards	Aerosmith	U2	Paul Simon (1.11)
12	Aerosmith	Michael Jackson	Bon Jovi	Eric Clapton (1.10)
13	Michael Jackson	Prince	Jeff Beck	Janet Jackson (1.0)
14	Prince	Tsuyoshi Nagabuchi	Prince	Rod Stewart (1.0)
15	U2	Bob Dylan	Mick Jagger	Bob Dylan (1.0)
16	Neil Young	Paul Simon	George Michael	George Michael (1.0)
17	Bryan Adams	Masaharu Fukuyama	Aerosmith	Mariah Carey (1.0)
18	Rod Stewart	Keith Richards	Stevie Wonder	Tina Turner (1.0)
19	Tomoyasu Hotei	U2	Wings	Bjork (1.0)
20	KinKi	Bryan Adams	Paul Simon	Richard (1.0)
	nDCG@20 = 0.808	nDCG@20 = 0.817	nDCG@20 = 0.879	nDCG@20 = 1.0

train's name, six hypernyms are related to a sea's name, and eight hypernyms are related to a song's name. In the proposed method, only the names of trains, such as "Twilight Express" and "Hatsukari," were included in top 50 coordinate terms because our method was profoundly affected by hypernyms that were related to a train's name. Average people will think that the names of seas, such as "the Pacific Ocean" and "Okhotsk Sea," are appropriate coordinate terms of "Japan Sea," and they do not know that "Japan Sea" could be related to the name of a train or a song. Thus, the appropriateness of names of trains and songs are low, and the proposed method does not achieve satisfying results. One approach to solve this problem is to cluster hypernyms based on n-gram similarities between hypernyms and the degree of duplication of their hyponyms, and to discover appropriate coordinate terms in each cluster using the cluster's hypernyms.

Another case is for a query that has few hypernyms. For example, the query "vending machine" had only two hypernyms, "sales method" and "business operator/distributor." In the proposed method, 49 terms had the same degree of coordination and were ranked first. This result defies our objective, which is to rank coordinate terms according to appropriateness. One ap-

Table 4.6: Ranking results of hypernyms for the proposed method for the query "Nintendo DS" (numbers in the parentheses indicate answer score).

rank	$\beta = 0$	$\beta = 0.3$
1	work	game device
2	appearance work	home computer game
3	game	peripheral device
4	game device	portable game device
5	home computer game	game hardware that uses ROM software
6	product	game software
7	game work	portable game device
8	biggest-selling computer game	Nintendo hardware
9	song content	computer software · game
10	Gundam series game	consumer game
nDCG@10	0.571	0.837

rank	$\beta = 1$	answer data
1	computer software · game	portable game device (2.00)
2	peripheral device	game device (1.91)
3	game software	computer game (2.00)
4	Nintendo hardware	Nintendo hardware (1.91)
5	game hardware that uses ROM software	game (1.91)
6	available terminal	home computer game (1.91)
7	brain training game	portable video game player (1.91)
8	goods · service	product (1.80)
9	portable video game player	Nintendo software (1.73)
10	portable game device	consumer game (1.70)
nDCG@10	0.703	1.0

proach to solve this problem is to combine the hypernym-hyponym dictionary used in this research with other dictionaries, such as WordNet [21, 45]. This would enable us to obtain more hypernyms and hyponyms, and to construct a larger bipartite graph.

4.4.2 Hypernyms

Table 4.6 presents results for an example for which the proposed method determined appropriate hypernyms with high accuracy. The table presents the top 10 hypernyms from the proposed method and the top 10 terms in terms of answer scores.

Table 4.6 presents the results of a query "Nintendo DS." When β was 0, hypernyms with many hyponyms, such as "work" and "product," were ranked higher. When β was 1, hypernyms labeled inappropriate because of the meaning being too narrow, such as "brain training game", were ranked higher. When β was 0.3, the results were well balanced and achieved the best nDCG value of 0.837.

4.5 Conclusion

In this chapter we have proposed methods for ranking coordinate terms and hypernyms of a query according to their appropriateness. The proposed method first creates a bipartite graph based on hypernyms of a query and hyponyms of each hypernym using a hypernym-hyponym dictionary. Subsequently, we applied a HITS-based algorithm to the graph and ranked coordinate terms and hypernyms. The experimental results using 50 queries indicate that the proposed method can rank appropriate coordinate terms and hypernyms higher than other comparable methods.

In the future, we will conduct more detailed experiments. Although we discarded terms that assessors did not understand, we plan to allow assessors to search the meanings of unknown terms and label their appropriateness, which will enable us to evaluate methods more accurately and will facilitate more in depth discussions.

In this chapter, we only targeted queries that occur in the titles of articles in the Japanese Wikipedia because we use a hypernym/hyponym extraction tool. Thus, in order to solve this problem, applying the proposed method to other data, such as WordNet [21, 45], would also be work of future interest.

CHAPTER 5

DISCOVERING UNEXPECTED INFORMATION BASED ON THE POPULARITY OF TERMS AND THE TYPICALITY OF RELATIONSHIPS BETWEEN TERMS

5.1 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 [64] has been proposed as a state-of-the-art text-based ranking function, and HITS [36] and PageRank [7] are link-based ranking algorithms. Based of these studies, many additional studies have reported improved methods for the retrieval of appropriate query results [18, 26, 28, 72].

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 [39, 42, 50, 54], although there has been a great deal of research focused on extracting unexpected or unusual frequent rules in the field of data mining [5, 57, 58, 75]. 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 "Hiromitsu Ochiai," it is well known that "Hiromitsu

[1]http://www.google.com
[2]http://www.yahoo.com
[3]http://www.bing.com

Ochiai was a leading hitter," however, it is generally unknown that "Hiromitsu Ochiai is a Gundam maniac." This information can be unexpected for users who know about "Ochiai Hiromitsu" and "Gundam" but do not know that there is a relationship. The user can find commonly 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 if top ranked web pages include unexpected information, it is usually buried in 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, unexpected information about a building or the surrounding area may be useful. Hence our objective is to discover unexpected information relative to keywords, such as specific people, facilities, and regions.

In this research, we target information that contains two objects. For example, in the information "Hiromitsu Ochiai was a leading hitter," one object is "Hiromitsu Ochiai" and the other is "leading hitter." We denote an object provided 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 5.2. Our approach involves the following three steps:

1. Given a query keyword (theme term) q, we collect its related terms $L_q = \{e_1, e_2, \cdots e_n\}$.

2. We compute the unexpectedness of each related term e_i for q on the basis of the typicality of the relationship between q and e_i, and the popularity of e_i.

3. We find information that includes an unexpected related term detected in step (2).

In step (1), we use Wikipedia[4] to collect a very large set of related terms. In step (2), we utilize the link structure between terms in Wikipedia and the super sub relation between terms. This step detects that, for example, "Gundam" has higher unexpectedness than "baseball" for a theme term "Hiromitsu Ochiai." We evaluate the unexpectedness of each related term e_i for q on the basis of relationships of the coordinate terms of q and e_i, and the popularity of e_i. In step (3), 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. One of the characteristics

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

of the proposed method is that we discover unexpected information using only the link structure and the super sub relation between terms obtained from Wikipedia.

In this research, we assume that the common perception of a theme term can be estimated by aggregating information from Wikipedia. Thus, information discovered by the proposed method is unexpected for ordinary people. Although our final goal is to discover useful unexpected information that attract users, we do not consider the usefulness of discovered unexpected information in this study

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

The remainder of this chapter is organized as follows. Section 5.2 explains the hypothesis of unexpected information as used in this research. Section 5.3 proposes methods for calculating the unexpectedness for each related term for a query. Section 5.4 describes the experimental set-up and reports results. A summary of this chapter and plans for future studies are presented in Section 5.5.

5.2 Unexpected Information

In this section, we explain the definition of unexpected information as used in this research.

We target information that contains two objects. Here an object is an essential element that constructs information when combined with another essential element (object). For example, in the case of the information "Hiromitsu Ochiai is a Gundam maniac," "Hiromitsu Ochiai" and "Gundam" are objects because they are important elements. This information could be shown when the user conducts a web search with the query "Hiromitsu Ochiai," or browses a news article about "Hiromitsu Ochiai." In these situations, we find unexpected information about the input keyword "Hiromitsu Ochiai." We denote the object 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 "Hiromitsu Ochiai," for example "leading hitter," "Akita Prefecture," and "Gundam," among many others.

When two terms have a common hypernym, they are coordinate terms. For example, "Hiromitsu Ochiai" and "Sadaharu Oh" are coordinate terms because they have a common hypernym, i.e., "baseball player." "Hiromitsu Ochiai" and "Taro Aso" are also coordinate terms because of the common hypernym, "human beings." However, "Sadaharu Oh" is a more appropriate coordinate term because "Hiromitsu Ochiai" and "Sadaharu Oh" have many common hypernyms in addition to "baseball player," such as "male" and "home run king." Conversely, "Taro Aso" is

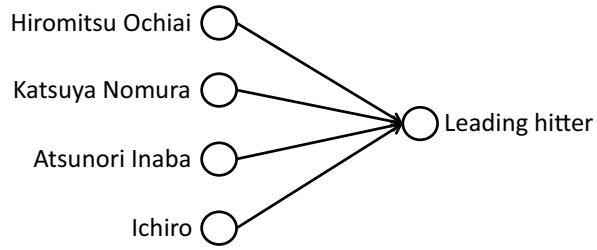

Figure 5.1: Related term "batting champion" is also related to appropriate coordinate terms of "Hiromitsu Ochiai."

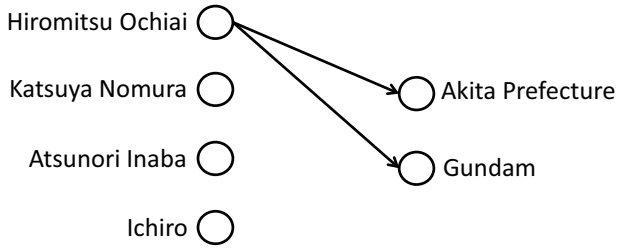

Figure 5.2: Related terms "Akita Prefecture" and "Gundam" are not related to appropriate coordinate terms of "Hiromitsu Ochiai."

a less-appropriate coordinate term. There are degrees of difference among the coordinate terms of a theme term. The appropriateness of coordinate terms for a term will be discussed later in Section 5.3.3.

To describe the type of information people perceive as unexpected relative to the theme term, its related term, and their coordinate terms, we examine four examples, each with the theme term "Michael Jackson." The information "Hiromitsu Ochiai was a leading hitter" is not unexpected to most people because it is well known that people who win batting titles are baseball players. In other words, appropriate coordinate terms of "Hiromitsu Ochiai" also have the term "leading hitter" as a related term (Figure 5.1). In this case, the relationship between "Hiromitsu Ochiai" and "leading hitter" is typical in *central tendency* because there are many relationships that are similar to the relationship between "Hiromitsu Ochiai" and "leading hitter."

For the information "Hiromitsu Ochiai is from Akita Prefecture" and "Hiromitsu Ochiai is a Gundam maniac," appropriate coordinate terms of "Hiromitsu Ochiai" may not have "Akita" or "Gundam" as related terms, as is shown in Figure 5.2. Although these two examples have the same structure, the information "Hiromitsu Ochiai is from Akita Prefecture" may not be common knowledge; however, it is not entirely unexpected information. All Japanese baseball players are from a certain prefecture; therefore, this information is just an example of that fact. That is, appropriate coordinate terms of "Hiromitsu Ochiai" have appropriate coordinate terms of "Akita

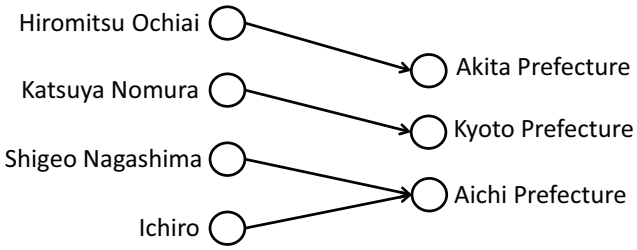

Figure 5.3: Appropriate coordinate terms of "Hiromitsu Ochiai" include appropriate coordinate terms of "Akita Prefecture" as a related term.

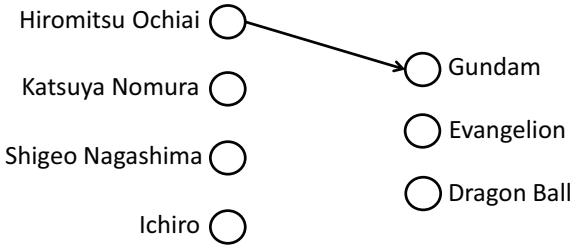

Figure 5.4: Appropriate coordinate terms of "Hiromitsu Ochiai" do not include appropriate coordinate terms of "Gundam" as a related term.

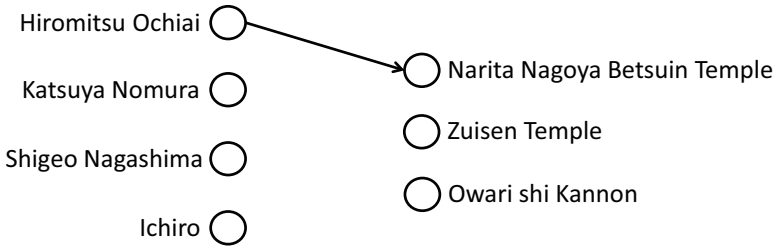

Figure 5.5: Appropriate coordinate terms of "Hiromitsu Ochiai" do not include appropriate coordinate terms of "Naritasan Nagoya Betsuin Daisyoji Temple" as a related term.

Prefecture" as their related terms (Figure 5.3). The relationship between "Hiromitsu Ochiai" and "Akita Prefecture" is also typical in *central tendency* because there are many relationships that are similar to the relationship. In contrast, most people do not expect baseball players to be an animaniac. Consequently, the degree of unexpectedness of the information "Hiromitsu Ochiai is a Gundam maniac," is quite high. In other words, appropriate coordinate terms of "Hiromitsu Ochiai" do not have appropriate coordinate terms of "Gundam" as their related terms (Figure 5.4). In this case, the relationship between "Hiromitsu Ochiai" and "Gundam" is atypical in *central tendency* because there are few relationships that are similar to that relationship.

We also consider the information "Hiromitsu Ochiai prayed for victory at Narita Nagoya Betsuin Temple." In this case, appropriate coordinate terms of "Hiromitsu Ochiai" do not have

appropriate coordinate terms of "Narita Nagoya Betsuin Temple" as their related terms (Figure 5.5). The relationship between "Hiromitsu Ochiai" and "Narita Nagoya Betsuin Temple" is also atypical in *central tendency* because there are few relationships that are similar to the relationship. However, the degree of unexpectedness of this information would be low because the term "Narita Nagoya Betsuin Temple" is not generally known; therefore, the popularity is low. We hypothesize that people do not perceive information as unexpected if it includes an unknown related term. Therefore, we must consider the popularity of each related term.

From the above explanation, we hypothesize that information is unexpected if it includes a related term that has an atypical relationship with the theme term and the popularity of the related term is high. Given a theme term q and its related term e, we define a function $f_{typ}(q, e)$ that represents the typicality between q and e. The function $f_{pop}(e)$ represents the popularity of e. We then define the following function f that combines these functions to calculate the unexpectedness of the pair of q and e.

$$f_{unexp}(q, e) = f(f_{typ}(q, e), f_{pop}(e)) \tag{5.1}$$

5.3 Methodology

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 typicality of a relationship $f_{typ}(q, e_i)$ between q and each related term.
4. Calculate the popularity $f_{pop}(e_i)$ for each related term.
5. Calculate the unexpectedness $f_{Unexp}(q, e_i)$ of each related term for q.

Figure 5.6 shows an overview of ranking unexpected related terms for the query "Hiromitsu Ochiai." We explain each step in detail in the following subsections.

5.3.1 Related Terms

In this paper, we regard anchor texts in a Wikipedia article of the theme term q as the related terms for q [5]. 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 is that

[5] We use the Japanese Wikipedia database dumped in July 2008.

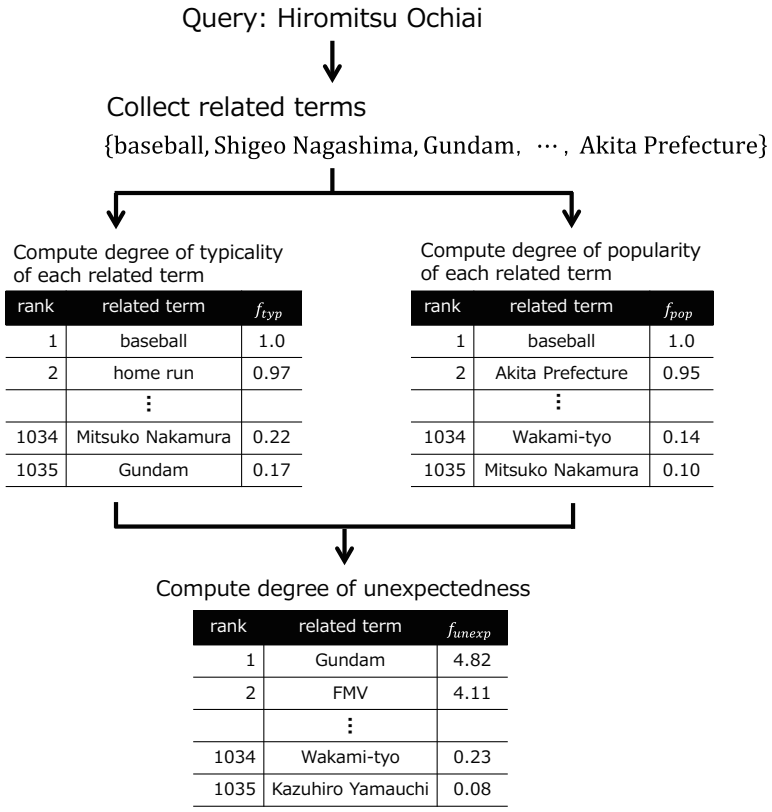

Figure 5.6: Overview of ranking unexpected related terms for the query "Hiromitsu Ochiai."

there are fewer noise terms in Wikipedia articles as they generally focus on information about a theme term q, whereas SERP would extract many noise terms from sentences that are unrelated to q. The second is that Wikipedia articles primarily contain objective information. We do not target unexpected information derived from personal opinions or impressions; we are only interested in information written from an objective perspective. And the final 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 [6]. Therefore, we regard linked terms in q's Wikipedia article as related terms of q. For all of these reasons, we collect all Wikipedia anchor texts in an article of q as related terms for q.

5.3.2 Hypernyms and Coordinate Terms

As discussed in Section 5.2, we identified unexpected information on the basis of a theme term, its coordinate terms, terms related to the subject term, and their coordinate terms. To collect coordinate terms, we used the hypernym/hyponym extraction tool used in Section 4.2.1. For instance, "Hiromitsu Ochiai" has a total of 45 hypernyms such as "baseball manager," "baseball player"

[6] http://ja.wikipedia.org/wiki/Wikipedia:記事どうしをつなぐ

and “human being.” If a term has at least one common hypernym with “Hiromitsu Ochiai,” the term is a coordinate term of “Hiromitsu Ochiai.”

5.3.3 Typicality of the Relationship between a Theme Term and its Related Term

Before explaining our proposed method in detail, we will describe it visually. In Figure 5.7, vertices represent terms and 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 $rel(t)$.

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

In Figure 5.7, the black circle, white circle, black triangle, and 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) indicates that there is an edge between a vertex n_1 and a vertex n_2.

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

This graph does not include edges between the theme term and its related terms because the objective is to demonstrate the ease of reaching all related terms of a theme term. We assume

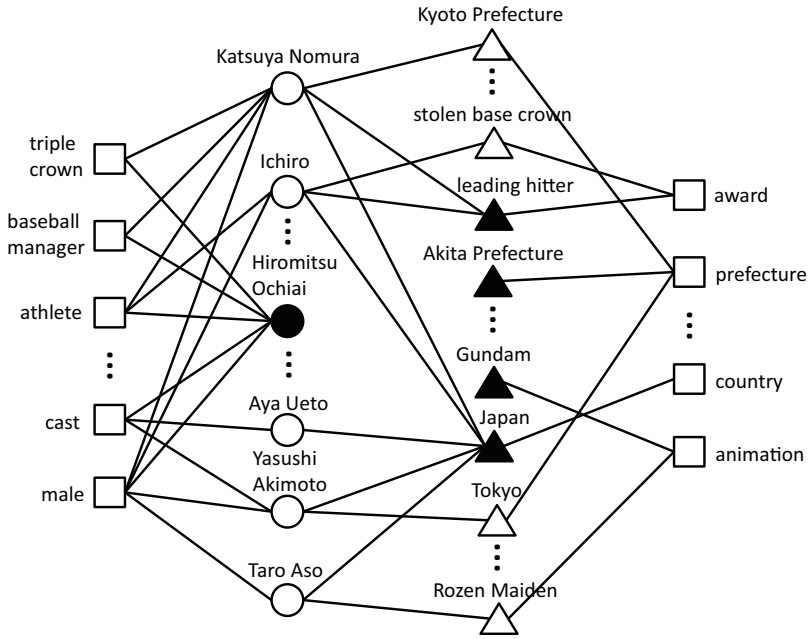

Figure 5.7: Example of the graph for a theme term "Hiromitsu Ochiai:" black circle vertex: a theme term; white circle vertex: a term in C_q; black triangle vertex: a term in L_q; white triangle vertex: a term in L_c; square vertex: a term in H_q or H_{lq}.

that if it is easy to reach a specific related term from a theme term, the related term is expected. As indicated previously, the term "batting title" is not unexpected for "Hiromitsu Ochiai." As is shown in Figure 5.7, there are many paths to reach "batting title" from "Hiromitsu Ochiai" through appropriate coordinate terms such as "Hiromitsu Ochiai $\rightarrow$ baseball player $\rightarrow$ Ichiro $\rightarrow$ batting title" and "Hiromitsu Ochiai $\rightarrow$ baseball manager $\rightarrow$ Katsuya Nomura $\rightarrow$ batting title." In this case, it is easy to reach the related term. In the case of "Akita Prefecture," there may be very few paths to reach "Akita Prefecture" directly in one step from appropriate coordinate terms of "Hiromitsu Ochiai." However, there are many paths to reach "Akita Prefecture" from appropriate coordinate terms through the hypernyms of "Akita Prefecture" such as "Hiromitsu Ochiai $\rightarrow$ baseball player $\rightarrow$ Ichiro $\rightarrow$ Aichi Prefecture $\rightarrow$ prefectures $\rightarrow$ Akita Prefecture" and "Hiromitsu Ochiai $\rightarrow$ baseball manager $\rightarrow$ Katsuya Nomura $\rightarrow$ Kyoto Prefecture $\rightarrow$ prefectures $\rightarrow$ Akita Prefecture." On the other hand, there are no paths to reach "Gundam" from appropriate coordinate terms, even through the hypernyms of "Gundam." There may be a few paths from less-appropriate coordinate terms directly or through the hypernyms; however, we assume that it is difficult to reach "Gundam" from "Hiromitsu Ochiai" and that there is potential for an unexpected term. For example "Hiromitsu Ochiai $\rightarrow$ male $\rightarrow$ Gackt $\rightarrow$ Gundam" and "Hiromitsu Ochiai $\rightarrow$ human beings $\rightarrow$ Taro Aso $\rightarrow$ Rozen Maiden $\rightarrow$ animation $\rightarrow$ Gundam."

To evaluate the degree of typicality 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 typicality of the relationship between them by considering the strength of the relationship as the ease of 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 typicality.

In the following subsections, we divide the graph into three subgraphs and evaluate the degree of typicality of the relationship between a theme term and each of its related terms.

Degree of Coordination to a Theme Term

To compute the degree of coordination to a theme term, we use the method based on the one proposed in Chapter 4. To apply the method, we first 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. In this chapter, the weight of the edge from h_i to c_j and from c_j to h_i is represented by $w_{ij}^{hc} = \frac{1}{|hypo(h_i)|}$ and $w_{ji}^{ch} = \frac{1}{|hyper(c_j)|}$, respectively. Then the SALSA algorithm [37] is applied to G_1. We denote the convergent scores of $c_j \in C_q$ as $f_{coordinate}\,(q, c_j)$, which represents the degree of coordination of c_j to q.

Typicality of Relationship between a Theme Term and each of its Related Terms on the basis of Coordinate Terms of the Theme Term

We calculate the degree of typicality of a relationship between a theme term and each of its related term on the basis of links from coordinate terms of the theme term to its related terms, links between related terms and links from related terms to coordinate terms of the theme term. First, we construct 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 $y \in C_q \cup L_q \cup L_c$ is a related term of $x \in C_q \cup L_q \cup L_c$, then there is an edge from x to y, and there could be edges from $x \in L_q$ to $y \in L_q$. The direction of an edge means that there is a link to t_j in a Wikipedia article where the title of an article is t_i. We assume that if t_i has a high degree of typicality to a theme term, t_j also has a high degree of typicality to the theme term because t_j is related to t_i. In other words, we assume that the degree of typicality to a theme term propagates according to the link structure. Our approach also has another advantage. If there is no link to t_i in a Wikipedia article where the title of an article is t_i, the degree of typicality of t_j to a theme term does not propagate to t_i. This is desirable because no link between t_i and t_j indicates that t_i is not related to t_j.

Our assumption to estimate the degree of typicality of a relationship between a theme term and each of its related terms has the following characteristics.

1. On graph G_2, a term has a typical relationship with a theme term if the term is linked to by many terms that have typical relationships with the theme term.

2. On graph G_2, appropriate coordinate terms of a theme term and a term that is linked to by appropriate coordinate terms of the theme term have typical relationships with the theme term.

To reflect the above characteristics, we use the biased PageRank [28] algorithm because this algorithm has the following two characteristics.

1. A web page is important if many other important web pages link to it.

2. A web page is important if web pages that are known to be important link to it before applying the biased PageRank algorithm.

We regard "important web pages" in the first characteristic of the biased PageRank algorithm as "terms that have typical relationships with the theme term," which corresponds to the first characteristic of our assumption. We regard "web pages that are known to be important" in the second characteristic of the biased PageRank algorithm as "terms that are known to have typical relationships with the theme term," which corresponds to the second characteristic of our assumption.

Before describing the detail of the biased PageRank algorithm, let us first describe the PageRank algorithm [7]. PageRank 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. This simple idea leads to the following equation:

$$r_{i+1}(v) = \alpha \sum_{u \in B_v} \frac{r_i(u)}{|F_u|} + \frac{1-\alpha}{N}. \tag{5.2}$$

Throughout this paper, we set α as 0.85, following the original PageRank algorithm. To evaluate the ease of reaching each related term from q, we revise Equation 5.2 on the basis of biased

PageRank:

$$r_{i+1}(v) = \alpha \sum_{u \in B_v} \frac{r_i(u)}{|F_u|} + (1-\alpha)\frac{f_{ini}(v)}{\sum_{t \in C_q} f_{ini}(t)}. \tag{5.3}$$

$f_{ini}(v)$ is the initial value of vertex v, which is defined as follows:

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

Here, $f_{co}(v)$ is the degree of coordination of term v to q. We apply this process to graph G_2. A vertex with a low score in L_q is a term that has an atypical relationship with q.

Typicality of Relationship between a Theme Term and each Related Term on the basis of Coordinate Terms of the Related Term

Finally, we evaluate the degree of typicality of a relationship between a theme term and each related term $e_i \in L_q$ by considering the coordinate terms of e_i. 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. Our assumption to estimate the degree of typicality of a relationship between a theme term and each of its related terms has the following four characteristics:

1. If many coordinate terms of a related term t have a typical relationship with a theme term, t also has a typical relation ship with the theme term.

2. If many coordinate terms of a related term t have an atypical relationship with a theme term, t also has an atypical relation ship with the theme term.

3. If a related term t has a typical relationship with a theme term after applying the biased PageRank algorithm, t has a typical relationship with the theme term regardless of t's coordinate terms.

4. If a related term t has an atypical relationship with a theme term after applying the biased PageRank algorithm, t has an atypical relationship with the theme term regardless of t's coordinate terms.

We apply the Co-HITS algorithm [18] to the bipartite graph for the following two reasons. First, we must obtain the coordinate terms of a related term in the above characteristics 1 and 2, and we use the SALSA algorithm to obtain coordinate terms of a theme term. The SALSA

algorithm is a special case of the Co-HITS algorithm; when we set $\lambda_u = \lambda_v = 1$ in Equations 5.4 and 5.5, the SALSA algorithm is equal to the Co-HITS algorithm.

To describe the second reason, let us consider a bipartite graph (V_1, V_2, E). Edges do not exist between vertices in V_1 and between vertices in V_2. Edges exist only between vertices in V_1 and V_2. The Co-HITS algorithm considers the initial value of each vertex. By changing a parameter, a vertex with a high (low) initial value can have a high (low) value even after applying the Co-HITS algorithm. This property is suitable for reflecting the characteristics 3 and 4. We regard "a vertex with a high initial value" as "a related term t that has a typical relationship with a theme term after applying the biased PageRank algorithm," and regard "a vertex with a low initial value" as "a related term t that has an atypical relationship with a theme term after applying the biased PageRank algorithm." In addition, the Co-HTIS algorithm has two parameters that are controlled individually. One is common to vertices in V_1, and the other is common to vertices in V_2. This property is also suitable for reflecting the characteristics 3 and 4. We do not need to consider the initial values of hypernyms of a related term because they do not have the degree of typicality as their initial values; however, we must consider the initial values of a related term and its coordinate terms.

The Co-HITS algorithm is described as follows. Let x_i and y_j denote the degree of authority of $u_i \in C_t$ and the degree of the hub of $v_j \in H_t$, respectively. We calculate the score of each vertex with the following equations:

$$x_i = (1 - \lambda_u)x_i^0 + \lambda_u \sum_{v_j \in H_t} w_{ji}^{vu} y_j, \tag{5.4}$$

$$y_j = (1 - \lambda_v)y_j^0 + \lambda_v \sum_{u_i \in C_t} w_{ij}^{uv} x_i, \tag{5.5}$$

where x_i^0 and y_j^0 represent the initial scores for terms u_i and v_j, respectively. The initial score of each hypernym in H_t is zero because the degree of importance of each hypernym is not predetermined. 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 the previous step. If a vertex in C_{e_i} is not included in graph G_2, its initial score is zero. 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 parameter λ_u in Section 5.4.2. Higher value for λ_u emphasizes the result of the biased PageRank algorithm for characteristics 3 and 4. In this case, the Co-HITS algorithm is equal to the personalized PageRank algorithm proposed by Taher *et al.* [73].

We conduct the operation for each related term of q. Let $f_{typ}(q, e_i)$ represent the score calculated by Equation 5.4.

5.3.4 Popularity of a Related Term

One way to calculate the degree of popularity of a term is by using the web hit count of the term. A term with a high hit count potentially infers the frequent use of that term. To obtain the web hit count of a term, we typically use a web search API provided by a web search engine. However, it is difficult to obtain the web hit count of a huge number of terms because there is a restriction on the pay-per-use of API and the web search API service was terminated by Yahoo! JAPAN on August 14, 2013 [7].

In this research, we use the PageRank [7] score of articles as the degree of popularity. In the PageRank algorithm, an article that is referenced by many good articles has a high PageRank score. We assume that the title of such an article is generally well known. Thus, we apply the PageRank algorithm to all articles in Wikipedia on the basis of the link structure. The degree of popularity 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 $f_{pop}(e_i)$.

5.3.5 Unexpectedness

Given the theme term q, we calculate the degree of typicality of a relationship $typ(q, e_i)$ between q and each of its related terms e_i. We have established that there is a higher degree of unexpectedness when there is a lower typicality value; therefore, we use the inverse of $typ(q, e_i)$. For the degree of popularity of each related term, a higher degree of popularity results in a higher degree of unexpectedness. Based on these ideas, the degree of unexpectedness $f_{unexp}(q, e_i)$ is calculated by the following equation:

$$f_{unexp}(q, e_i) = \frac{1}{f_{typ}(q, e_i)} \cdot f_{pop}(e_i). \tag{5.6}$$

5.4 Experiments

We conducted two related experiments to examine the effectiveness of the proposed method:

1. Experiment for term popularity determination.

2. Experiment for discovering unexpected information.

We created a query set consisting of 75 theme terms in the following five categories: names of people, facilities, regions, products, and organizations. These queries were used in experiment 2. Each category included 15 theme terms. If a user is not familiar with the theme term, all information will not be unexpected for that user. Thus, we selected terms that appear in the top

[7] http://techblog.yahoo.co.jp/topics/search_api_close/

Table 5.1: Examples of experimental queries.

Category	Query with more than 250 related terms	Query with fewer than 250 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	Mitsui Group, University cooperative

5% of PageRank scores among all Wikipedia articles. The number of articles was $17,325$. We assumed that the fewer the number of related terms, the lower 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 250 related terms, and group (b) included articles that had less than 250 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 articles in each category were randomly selected from group(b). We used the title of each article as a query, i.e., a theme term. Examples of the query set are shown in Table 5.1.

5.4.1 Term Popularity Determination

In this section, we evaluate the method proposed in Section 5.3.4 for calculating the degree of popularity of a term.

First, the scores of all terms on Wikipedia were computed using the method described in Section 5.3.4. Next, all terms were divided into 10 blocks according to their scores, and 10 terms were randomly sampled from each block. We used a total of 100 terms for the evaluation. Three males in their twenties evaluated the degree of popularity independently [8]. We first showed a query and asked assessors to label each of its coordinate terms on a scale of 1-5 from unpopular to popular. Then we calculated the average degree of popularity for each term and used it as answer data.

The comparative method regarded the web hit count of a term as the degree of popularity. We used the Bing Search API [9] to obtain the web hit count.

Table 5.2 shows the kappa agreement with quadratic weight [22] among the assessors. Significance test results showed that all scores in Table 5.2 were statistically significant at $\alpha = 0.01$. Table 5.3 shows the Pearson correlation between the answer data and the comparative method or our proposed method. The proposed method achieved 0.834, which indicates a significant correlation at $\alpha = 0.01$, and outperformed the comparative method. From these results, we can

[8] Although one assessor is an author of this paper, the experimental condition were the same for all assessors.
[9] `http://datamarket.azure.com/dataset/bing/search`

Table 5.2: Kappa agreement of popularity scores between assessors.

	assessors 1 and 2	assessors 2 and 3	assessors 3 and 1
κ agreement	0.775	0.868	0.830

Table 5.3: Pearson's product-moment correlation coefficient between the popularity calculated by a baseline method or our proposed method and the popularity determined by assessors.

	comparative method	proposed method
Pearson correlation	0.816	0.834

conclude that the proposed method can accurately estimate the degree of popularity of a term.

5.4.2 Unexpected Information Discovery

The objective of this experiment was to clarify two research questions:

- Is considering the degree of popularity of related terms important to the discovery of unexpected information?

- 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 three proposed methods and compared them with four simpler methods. The three proposed methods calculate the degree of unexpectedness of each related term using Equation 5.6. To compare the impact of λ_u in Equation 5.4, we set λ_u to 0.25, 0.5, and 0.75. A method using Equation 5.6 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 use three additional simple methods to answer the first research question. In these methods, only the degree of the relationship between a theme term and a related term is evaluated, and the degree of popularity of related terms is neglected. The unexpectedness score of related term e_i for the theme term q is calculated by

$$f_{unexp}(q, e_i) = \frac{1}{f_{typ}(q, e_i)}. \tag{5.7}$$

In these methods, we also set λ_u to 0.25, 0.5, and 0.75. We denote each method as TYP_{25}, TYP_{50} and TYP_{75}.

We also proposed a simple method to answer the second research question. In this method, we get the web hit count for each pair of (q, e_i) [10]. The query is "q $\wedge e_i$" for the pair of (q, e_i). Then, the related terms are ranked in ascending order of hit count. That is, we assume that if a related term has low a co-occurrence frequency with q, the term is unexpected for q. We denote this method as HIT.

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 from the article that includes the related term. If the related term is included in more than one sentence, we extract the first sentence that uses the term.

In the remainder of this section, we describe the experimental procedure, present metrics for evaluation, and discuss the results.

Procedure

In this experiment, we recruited five assessors and administered a questionnaire. Two males and one female were in their thirties and two females were in their twenties [11]. We created the questionnaire as follows. First, a theme term was used with each method. Given a theme term, each of the seven 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 assessors to label each pair of related terms 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; each questionnaire corresponded to a single theme term. We ordered five sets of questionnaires taking the order effect into consideration. Five assessors 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" as highly unexpected and output the corresponding information "Kimiko Date is now living in Monaco." The five assessors labeled the unexpectedness of this information as 4, 3, 2, 3, and 2. The average degree of unexpectedness was 2.8.

Metrics for Evaluation

We used nDCG [32] and the Normalized Weighted Reciprocal Rank(NWRR) [67] as evaluation metrics.

When we present information to the user, the number of terms is limited and it is preferable

[10] In this research, we used the Yahoo! Web Search API (`http://developer.yahoo.co.jp/webapi/search/websearch/v1/websearch.html`) before Yahoo! JAPAN terminated the service.

[11] Authors are not included in the assessors

Table 5.4: Kappa agreement of unexpectedness scores between assessors. $**$ represents that inter-assessor agreement was statistically significant at $\alpha = 0.01$.

	assessor 1	assessor 2	assessor 3	assessor 4
assessor 2	0.210^{**}			
assessor 3	0.264^{**}	0.0422		
assessor 4	0.437^{**}	0.164^{**}	0.331^{**}	
assessor 5	0.208^{**}	0.0462	0.206^{**}	0.268^{**}

to show at least one or more unexpected information items at a higher rank. NWRR can be regarded as a graded-relevance version of an RP. In this metric, the smallest penalty value is assigned to a theme term that is included in highly unexpected information. Let l_t denote the average evaluated score of the unexpectedness degree of information that includes the related term t. We used $L_t = 5 - l_t$ as the penalty value. In our experiment, each piece of information has an unexpected score judged by assessors. This score ranges from 1-4, and we regarded only the related term t with $L_t \geq 2.5$ as relevant unexpected information. Then WRR is calculated as follows:

$$WRR = \frac{1}{r_1 - 1/L_t}. \tag{5.8}$$

To normalize WRR, NWRR is defined as follows:

$$NWRR = \frac{1 - 1/L_{t_{max}}}{r_1 - 1/L_t}, \tag{5.9}$$

where t_{max} is the related term that obtained the highest unexpectedness degree. For each method, we calculated the average NWRR of theme terms. In this metric, the score is 1 if a method can place 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 the evaluators, the NWRR score is 0.

Results

Table 5.4 shows the kappa agreement with quadratic weight [22] among the assessors. Except for assessor 2, all scores in Table 5.4 were statistically significant at $\alpha = 0.01$. The low kappa agreement between assessors indicates that the unexpectedness of information highly depends on the assessor. Proposing a method specialized for each user to discover unexpected information would be interesting future work.

The nDCG scores for each method and category are shown in Table 5.5. In all categories, one of the three proposed methods resulted in the highest nDCG. TYP_{25}, TYP_{50}, and TYP_{75}

Table 5.5: Performance comparison of each category for seven methods measured by nDCG@5. $*$ ($\alpha = 0.05$) and $**$ ($\alpha = 0.01$) indidate significant differences with HIT.

method	person	region	product	facility	organization	average
HIT	0.705	0.757	0.773	0.787	0.780	0.760
TYP_{25}	0.805^{*}	0.792	0.837	0.800	0.853^{*}	0.817^{**}
TYP_{50}	0.807^{*}	0.803	0.839	0.800	0.857^{**}	0.821^{**}
TYP_{75}	0.807^{*}	0.808	0.841	0.804	0.852^{*}	0.822^{**}
$\mathbf{PR_{25}}$	$\mathbf{0.828}^{**}$	0.830	0.846^{*}	**0.825**	$\mathbf{0.860}^{**}$	$\mathbf{0.838}^{**}$
$\mathbf{PR_{50}}$	0.824^{**}	0.830	0.851^{*}	0.821	$\mathbf{0.860}^{**}$	0.837^{**}
$\mathbf{PR_{75}}$	0.818^{**}	$\mathbf{0.836}^{*}$	$\mathbf{0.858}^{**}$	0.820	0.854^{**}	0.837^{**}

Table 5.6: Performance comparison of each category for seven methods measured by NWRR.

method	person	region	product	facility	organization	average
HIT	0.307	0	**0.478**	0	0	0.157
TYP_{25}	**0.513**	0.118	0.319	0.215	0.165	0.266
TYP_{50}	0.506	0.118	0.327	0.199	0.177	0.266
TYP_{75}	0.506	0.163	0.332	0.194	0.177	0.274
$\mathbf{PR_{25}}$	0.434	0.184	0.341	**0.418**	**0.194**	**0.314**
$\mathbf{PR_{50}}$	0.418	0.184	0.361	0.241	**0.194**	0.280
$\mathbf{PR_{75}}$	0.421	**0.199**	0.361	0.241	**0.194**	0.283

followed those three methods. The results show that it is important to consider the degree of popularity of related terms to discover unexpected information. The HIT method returned the lowest scores in all categories. In the HIT method, terms with a low degree of popularity were often ranked high. There are two types of terms with low degree of popularity. The one is a term that is not related to a common topic for the theme term and its coordinate terms. In this case, our proposed method outperformed the HIT method because our proposed method considered the degree of popularity of a related term. The other type of term with low degree of popularity is related to a common topic for the theme term and its coordinate terms. In this case, our proposed method outperformed the HIT method not only because the proposed method considered the degree of popularity of a related term but also because it regarded a related term that has a relationship with many coordinate terms of the theme term as not an unexpected term. There was not a significant difference for the λ_u parameter in the Co-HITS algorithm.

The NWRR scores for each method in each category are shown in Table 5.6. On average, PR_{25} could discover more unexpected information at a higher rank than other methods. HIT and TYP_{25} obtained the highest scores in the product and person categories, respectively. However, the average scores for these two methods were lower than our three 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. In addition, the proposed methods could discover unexpected information in any category. As before, there was not a significant difference for the λ_u parameter in the Co-HITS algorithm.

We show some examples of information evaluated as unexpected information in Table 5.7. 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 death rate from lifestyle-related diseases such as a stroke." was discovered. In our method, other Japanese prefectures 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. "Nobita Nobi" is a cartoon character, and a related term "first-degree equation" was discovered as an unexpected term because most cartoon characters that are appropriate terms of "Nobita Nobi" are not related to any equation. However, in this case, we could surmise that this information was evaluated as unexpected not only because of the above reason but also because the character is famous for not being good at his studies. One challenge for the future is to consider the property of a theme term when we compute the degree of unexpectedness.

Table 5.7: Examples of discovered unexpected information.

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 death rate 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.

Table 5.8: Number and ratio of theme terms that could find unexpected information.

Category	Over 250	Under 250	Total
person	6/10	3/5	9/15
region	3/10	0/5	3/15
product	5/10	2/5	7/15
facility	4/10	0/5	4/15
organization	2/10	1/5	3/15

Finally, in Table 5.8, we show the number and ratio of theme terms in which we could discover at least one piece of unexpected information. On average, we could discover unexpected information in 40% of theme terms that had more than 250 related terms and in 24% of theme terms that had less than 250 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 included 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 "digital camera" article 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;" however, it is related to many other electrical products that are appropriate coordinate terms of "digital camera." Therefore, our method could not discover this information.

5.5 Summary

In this paper, we proposed a new method for the discovery of unexpected information. In particular, we focused on two aspects: (1) the typicality of the relationship between a theme term and its related term, and (2) the popularity of each related term. We conducted an experiment to clarify the importance of considering these two aspects. Our results showed that the popularity of a related term was highly relevant to the 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 need to address the problem of removing noise terms. In addition, we need to consider

the credibility of unexpected information, especially when unexpected information is discovered from more general web pages. False or untrue information is not useful. One method to verify credibility is to assess the publisher. 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.

CHAPTER 6

DISCOVERING AN UNEXPECTED RELATIONSHIP BY MEASURING PERCEIVED STRENGTH OF THE RELATIONSHIP BETWEEN TERMS

6.1 Introduction

In Chapter5, we proposed methods for discovering unexpected information for an input query. Given an input query (object) of "Hiromitsu Ochiai," the proposed method first detected a term (attribute) "Gundam" as an unexpected term and then discovered unexpected information, i.e., "Hiromitsu Ochiai is a Gundam maniac." We calculated the unexpectedness between an object and an attribute based on the relations between the appropriate coordinate terms of an object and an attribute. For example, appropriate coordinate terms of "Hiromitsu Ochiai" such as "Shigeo Nagashima," "Katsuya Nomura," and "Sadaharu Oh" also have a relation with "leading hitter;" therefore, the relation between "Hiromitsu Ochiai" and "leading hitter" is popular or not unexpected. On the other hand, the appropriate coordinate terms of "Hiromitsu Ochiai" do not have a relation with "Gundam" and its appropriate coordinate terms such as "Evangalion" and "Dragon Ball;" therefore, the relation between "Hiromitsu Ochiai" and "Gundam" is unexpected.

However, even if the appropriate coordinate terms of an object have relations with an attribute, the relation between the object and the attribute is not always popular. For example, consider the relation between an object "Sanjusangen-do Temple" and an attribute "hipped roof." Although appropriate coordinate terms of "Sanjusangen-do Temple," such as "Daigo-ji Temple," "Ninna-ji Temple," and "To-ji Temple," also have relations with "hipped roof," the relation between

"Sanjusangen-do Temple" and "hipped roof" is not well-known. Therefore, it is evident that another method is required to calculate the perceived strength of a relation between terms.

Numerous studies have proposed to compute the strength of a relation between terms [6, 13, 24, 40, 46, 66, 71]. In these studies, the strength of a relation is computed by link structures on Wikipedia [46] and the co-occurrence frequency between terms on the Web [6, 40]. However, high strength of a relation between terms computed by such data does not guarantee high perceived strength of the relation.

If we can compute the perceived strength of a relation between terms, we can discover the following information:

(1) A relation with low perceived strength that has high strength on the Web.

(2) A relation with high perceived strength that has low strength on the Web.

This information can be unexpected because it is counter to people's expectations. In this chapter, we propose methods for computing the perceived strength of the relation between an object and an attribute on the basis of (i) the popularity of the object and (ii) the perceived strength of a relation between objects similar to the object and the attribute.

We conduct an experiment regarding the estimation accuracy of the perceived strength of a relation between an object and an attribute. We use 25 attributes in five categories: country, vegetable, tourist spot in Kyoto, electronic company, and baseball player. Our results demonstrate the effectiveness of considering the popularity of the object and the perceived strength of a relation between objects similar to the object and the attribute. In addition, we evaluate whether the above mentioned information, (1) and (2), are unexpected for assessors. Finally, given a pair of an object and an attribute, such as "India" and "coffee," we evaluate the prediction accuracy of the relation's unexpectedness on the basis of the strength of the relation on the Web and its perceived strength.

The remainder of this chapter is organized as follows. Section 6.2 explains the approach for estimating the perceived strength of a relation between an object and an attribute. Section 6.3 proposes methods for calculating the perceived strength of a relation. Section 6.4 describes the experimental results of the proposed methods and discusses the estimation of unexpectedness. A summary of this chapter and plans for future studies are presented in Section 6.5.

6.2 Approach

In this section, we describe the factors that affect the perceived strength of a relation between terms. Given a term pair (t_1, t_2), our goal is to compute the perceived strength of the relation between t_1 and t_2. To achieve this goal, given a category c, a set of terms $T_c = \{t_{o_1}, t_{o_2}, \cdots, t_{o_n}\}$

that belong to c, and a term t_a, we aim to compute the perceived strength of the relation between $t_{o_i} \in T_c$ and t_a, and rank $t_{o_i} \in T_c$ in descending order of scores. For example, when c is "country," T_c is a set of country names, and t_a is "wine," our proposed system returns the list of countries ranked in descending order of the perceived strength of the relation with wine. Hereafter, we denote $t_{o_i} \in T_c$ as an *object* and t_a as an *attribute*.

As discussed in Section 6.1, the strength of a relation between terms computed by data on the Web does not necessarily correspond to the perceived strength of the relation. In the following subsections, we propose two factors that affect the perceived strength of the relation between terms.

6.2.1 Popularity of an Object

Given a category "baseball player" and an attribute "Golden Glove Award," consider the perceived strength of the relation between a baseball player and the Golden Glove Award. First, consider that an object is "Hiromitsu Ochiai" and an attribute is "Golden Glove Award." In fact, the strength of the relation between "Hiromitsu Ochiai" and "Golden Glove Award" is low because Ochiai has not won the award. However, people think that "Hiromitsu Ochiai is highly relevant to Golden Glove Award because he is a famous baseball player," and they estimate the strength of the relation to be high. Next, consider that an object is "Hiromi Matsunaga." The strength of the relation between "Hiromi Matsunaga" and "Golden Glove Award" is high because Matsunaga has won the award several times. However, people think that "Hiromi Matsunaga is not relevant to Golden Glove Award because he is not a famous baseball player," and they estimate the strength of the relation to be low.

On the basis of these ideas, we formulate the following hypothesis.

HYPOTHESIS 1: If the popularity of an object is high (low), the perceived strength of the relation between the object and an attribute is estimated to be high (low).

6.2.2 Perceived Strength of a Relation between Similar Objects of an Object and an Attribute

Given a category "country" and an attribute "coffee," we discuss the perceived strength of the relation between a country and coffee. First, consider that an object is "Argentina" and an attribute is "coffee." In fact, the strength of the relation between "Argentina" and "coffee" is low because the amount of coffee consumed or produced in Argentina is not high. However, people think that "Argentina is highly relevant to coffee because we know that countries that are similar to Argentina such as Brazil, Colombia, and Mexico are highly related to coffee," and they estimate the strength of the relation between Argentina and coffee to be high. Next, consider that an object is "India." The strength of the relation between "India" and "coffee" is high because

coffee production is high in India. However, people think that "India is not relevant to coffee because they know that countries that are similar to India such as China, Pakistan, and Sri Lanka are not highly related to coffee," and they estimate the strength of the relation between India and coffee to be low.

On the basis of these ideas, we formulate the following hypothesis.

HYPOTHESIS 2: If the perceived strength of relations between objects that are similar to an object and an attribute is high (low), the strength of the relation between the object and an attribute is estimated to be high (low).

According to this hypothesis, the perceived strength of relations between countries that are similar to Colombia such as Argentina and Mexico and coffee also affects the perceived strength of the relation between "Colombia" and "coffee." This indicates that the perceived strength of the relation between an object and an attribute is recursively determined.

6.3 Methodology

In this section, we propose methods for computing the perceived strength of the relation between terms based on the two hypotheses given in Section 6.2.

We compute the perceived strength of a relation by extending methods that were proposed in previous studies. In Section 6.3.1, we explain the existing methods used in this chapter. Section 6.3.2 describes a method that considers the Hypothesis 1, and Section 6.3.3 describes a method that considers the Hypothesis 2. Finally, Section 6.3.4 describes a method that considers both the Hypothesis 1 and 2.

6.3.1 Existing Methods

In this chapter, the following three methods are used to compute the strength of the relation between terms.

WLM Method

The first method considered was proposed by Milne *et al.* [46]. This method computes the strength of the relation between a term t_1 and a term t_2 on the basis of the similarity between pages that t_1 links and those that t_2 links on Wikipedia. The similarity is computed by a TF-IDF-like method. Let s be a referrer page and t denote a page that s links. The weight of the link from s to t is given by:

$$w(s \to t) = \log\left(\frac{|W|}{|B_t|}\right) \text{ if } s \in B_t, \;\; 0 \text{ otherwise}, \tag{6.1}$$

where B_t is a set of t_1's referrer pages and W is all the pages on Wikipedia. We create a vector for each of t_1 and t_2 using Equation 6.1. That is, in Equation 6.1, s is t_1 or t_2. Let F_{t_1} be a set of

pages that t_1 links and F_{t_2} be a set of pages that t_2 links. t corresponds to each page in $F_{t_1} \cup F_{t_2}$. Finally, we compute the cosine similarity, denoted by $sim_{forward}(t_1, t_2)$, between the vector of t_1 and t_2.

This method also uses the similarity between referrer pages of t_1 and t_2. The similarity is computed by:

$$sim_{backward}(t_1, t_2) = \frac{\log\left(\max\left(|B_{t_1}|, |B_{t_2}|\right)\right) - \log\left(|B_{t_1} \cap B_{t_2}|\right)}{\log\left(|W|\right) - \log\left(\min\left(|B_{t_1}|, |B_{t_2}|\right)\right)}. \tag{6.2}$$

Finally, the strength of the relation between t_1 and t_2, denoted by $wlm(t_1, t_2)$, is given by:

$$wlm(t_1, t_2) = \frac{1}{2} \cdot sim_{forward}(t_1, t_2) + \frac{1}{2} \cdot sim_{backward}(t_1, t_2). \tag{6.3}$$

WebPMI Method

The second method is WebPMI [6, 40], in which the degree of relation between a term t_1 and a term t_2 is computed on the basis of their co-occurrence frequency:

$$web\_pmi(t_1, t_2) = \begin{cases} 0 & \text{if } hit(t_1, t_2) \leq c \\ \log_2\left(\frac{hit(t_1 \wedge t_2)/N}{(hit(t_1)/N) \times (hit(t_2)/N)}\right) & \text{otherwise,} \end{cases} \tag{6.4}$$

where $hit(t)$ is the Web hit count of t. We use the ClueWeb09 Japanese Dataset [1] to obtain the Web hit count. We used $c = 5$, according to Bollegala *et al.* [6]. N is the total number of Web pages, and $N = 67,337,717$ in the ClueWeb09 Japanese Dataset.

WebJaccard, WebDice, WebOverlap, and NGD also compute the strength of the relation between terms on the basis of the Web hit count. Among them, WebPMI has the highest accuracy [6, 40].

Noda Method

The third method proposed by Noda *et al.* [53] also computes the strength of the relation between two terms on the basis of their co-occurrence frequency. This method computes the strength of the relation between a term t_1 and a term t_2 as follows:

$$noda(t_1, t_2) = \frac{hit(t_1 \wedge t_2)^2}{hit(t_1) \cdot hit(t_2)}. \tag{6.5}$$

In this method, co-occurrence frequency of the two terms has greater impact on the strength of the relation than WebPMI.

[1] `http://lemurproject.org/clueweb09/`

6.3.2 Perceived Strength of a Relation on the basis of an Object's Popularity

On the basis of the Hypethesis 1 in Section 6.2.1, we compute the perceived strength of the relation between an object t_{o_i} and an attribute t_a by considering the popularity of t_{o_i}, which can be expressed as follows:

$$rel\_pop(t_{o_i}, t_a) = rel(t_{o_i}, t_a) \cdot pop(t_{o_i}), \tag{6.6}$$

where $rel(t_{o_i}, t_a)$ is the strength of the relation between t_{o_i} and t_a, which is computed by one of the three methods introduced in Section 6.3.1. $pop(t_{o_i})$ is the popularity of t_{o_i}, which is the logarithm of the number of Wikipedia articles that link to t_{o_i}.

6.3.3 Perceived Strength of a Relation on the basis of Similar Objects

Similar Objects

We regard appropriate coordinate terms of an object t_{o_i} as objects similar to t_{o_i}. A coordinate term of a term t is a one that has one or more hypernyms in common with t [55]. We have proposed a method for collecting appropriate coordinate terms for a given term. The details of this method can be found in Chapter 4.

Perceived Strength of a Relation

As mentioned in Section 6.2.2, when we compute the perceived strength of the relation between an object and an attribute on the basis of Hypothesis 2, it is necessary to recursively compute the value. Hence, we use the biased PageRank algorithm [28] because it has the following characteristics:

1. A Web page is important if several other important Web pages link to it.

2. A Web page is important if it is linked by Web pages that are known to be important before applying the biased PageRank algorithm.

Characteristic 1 corresponds to "the perceived strength of a relation between an object and an attribute is high if its appropriate coordinate terms have high perceived strength of relations with the attribute" in Hypothesis 2. Characteristic 2 reflects the strength of the relation between an object and an attribute computed by existing methods.

To apply the biased PageRank algorithm, we first create an adjacent matrix. Given $T_c = \{t_{o_1}, t_{o_2}, \cdots, t_{o_n}\}$, we create an n-dimensional square matrix in which the value of (i, j) element

is $f_{coordinate}(t_{o_j}, t_{o_i})$, which is computed by using a method described in Section 4.2.3. An adjacent matrix is created by normalizing the n-dimensional square matrix. The value of the (i, j) element of the adjacent matrix, denoted by $a_{i,j}$, is given by:

$$a_{i,j} = \frac{f_{coordinate}(t_{o_j}, t_{o_i})}{\sum_{t_{o_k} \in T_c} f_{coordinate}(t_{o_j}, t_{o_k})}. \tag{6.7}$$

Using this adjacent matrix, the perceived strength of the relation, denoted by $rel\_crd(t_{o_i}, t_a)$, between an object t_{o_i} and an attribute t_a is computed as follows:

$$rel\_crd_{l+1}(t_{o_i}, t_a) = \alpha \cdot \sum_{1 \leq j \leq n} a_{i,j} \cdot rel\_crd_l(t_{o_i}, t_a) + (1 - \alpha) \cdot \frac{rel(t_{o_i}, t_a)}{\sum_{t_{o_j} \in T_c} rel(t_{o_j}, t_a)}. \tag{6.8}$$

In Equation 6.8, the first and the second term correspond to the characteristic 1 and 2 in the biased PageRank algorithm, respectively. α is a damping factor, and ranges between $0 \leq \alpha \leq 1$. As the value of α increases, the impact of the perceived strength of relations between coordinate terms and an attribute also increases. We discuss the effectiveness of parameter α in Section 6.4.

6.3.4 Perceived Strength of a Relation on the basis of an Object's Popularity and Similar Objects

Finally, we compute the perceived strength of the relation between an object t_{o_i} and an attribute t_a on the basis of the popularity of t_{o_i} and objects similar to t_{o_i}. In this method, the initial value of the strength of the relation between t_{o_i} and t_a in the biased PageRank algorithm is replaced by the value computed in Section 6.3.2:

$$rel\_pop\_crd_{l+1}(t_{o_i}, t_a) = \alpha \cdot \sum_{1 \leq j \leq n} a_{i,j} \cdot rel\_pop\_crd_l(t_{o_i}, t_a) + (1 - \alpha) \cdot \frac{rel_{pop}(t_{o_i}, t_a)}{\sum_{t_{o_j} \in T_c} rel_{pop}(t_{o_j}, t_a)}. \tag{6.9}$$

6.4 Experiments

We conducted experiments to examine the effectiveness of our proposed method. The objective of our experiments was to verify the following research questions.

1. When the strength of a relation between terms is high (low) in data on the Web but the perceived strength of the relation is low (high), is the relation unexpected?

2. Is consideration of the popularity of an object and similar objects of the object important in computation of the perceived strength of the relation between the object and an attribute?

3. Given a pair of terms (i.e., object and attribute), can we estimate its unexpectedness?

Table 6.1: Categories, number of objects in each category, and attributes used in our experiment.

category	♯ object	attribute
country	169	wine, beer, coffee, pizza, banana
vegetable	117	vitamin C, dietary fiber, iron, protein, calcium
tourist spot in Kyoto	155	Nobunaga Oda, Hideyoshi Toyotomi, Ieyasu Tokugawa, Yoshimitsu Ashikaga, Rikyu Sen
electronics company	481	mobile phone, liquid crystal television, digital camera, refrigerator, personal computer
baseball player	157	home run king, leading hitter, stolen base crown, MVP, Golden Glove Award

Table 6.2: Number of objects in each category used for evaluation of degree of recognition of a relation.

category	♯ object
country	40
vegetable	39
tourist spot in Kyoto	13
electronics company	30
baseball player	32

6.4.1 Data set

We used the following five categories: country, vegetable, tourist spot in Kyoto, electronics company, and baseball player, and five attributes for each category. The author manually created object sets for each category. We selected objects and attributes that have Wikipedia articles in order to use the WLM method described in Section 6.3.1 and obtained an object's coordinate terms using the method described in Section 6.3.3. The number of objects in each category and their attributes are shown in Table 6.1. We used the Japanese Wikipedia database dumped in July 2008, where the value of $|W|$ in Equation 6.1 is $1,342,098$.

6.4.2 Questionnaire

The following two kinds of data are required to answer our three research questions.

- Data regarding the perceived strength of a relation between an object and an attribute.
- Data regarding the unexpectedness of a relation between an object and an attribute.

To collect the data, we recruited assessors through Lancers [2], which is a popular crowd sourcing marketplace in Japan. Typically, it is difficult for assessors to determine the strength of the relation and the unexpectedness between an object and an attribute if the object is unpopular. Thus, for each category, we first selected the top 40 objects in terms of referrer pages in Wikipedia. Then, we asked 10 assessors if they were familiar with each object. Objects that were known by more than five assessors were used in our experiments. Table 6.2 shows the number of objects that were known by more than five assessors for each category.

We created the questionnaire as follows. First, we computed the strength of the relation between each object in T_c and t_a based on the co-occurrence frequency on the Web. Next, scores were normalized such that the minimum value was 0 and the maximum value was 100. This process was conducted for all categories and attributes, and the pairs (c, t_a, $t_{o_i} \in T_c$, and the strength of the relation between t_a and t_{o_i}) were pooled. This gave us 770 pairs. We randomly selected 10 pairs and created a questionnaire; we created a total of $770/10 = 77$ questionnaires. In the questionnaire, we first showed assessors each pair and asked them to label the strength of the relation of each pair on a scale of 1-5 (very weak to very strong). Then, we showed the assessors the strength of the relation of each pair computed using the Web data and asked them to label the pair on a scale of 1-5 (expected to unexpected). Figure 6.1 shows the interface for the questionnaire. In the questionnaire, the following instructions were provided to assess unexpectedness.

- The number following each item indicates how often the two terms are discussed on the Web.
- "100" implies a strong relation and "0" implies no relation.
- When you cannot judge the degree of unexpectedness, choose "neither."

Each pair was labeled by 20 assessors.

A total of 157 assessors answered the questionnaires, and on an average, one assessor labeled 90.1 pairs. We calculated the average strength of the relation for each object and attribute pair. The average strength was regarded as the perceived strength of the relation between the object and attribute. Similarly, we computed the average assessor scores of unexpectedness for each pair and regarded this score as the unexpectedness score. For example, the average unexpectedness score for the pair "wine" and "Viet Nam" was 3.7.

[2] http://www.lancers.jp/

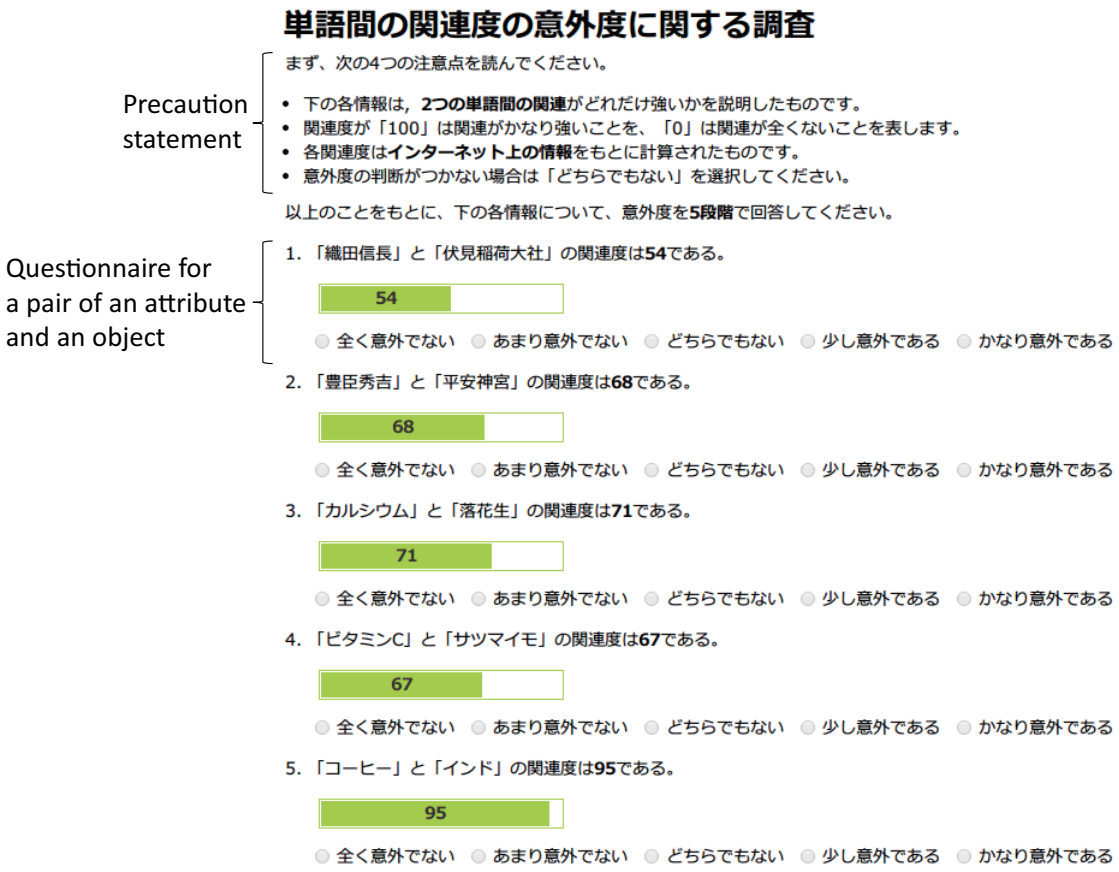

Figure 6.1: Interface used in the experiment.

6.4.3 Analysis of Unexpected Information

In this section, we answer research question 1: when the strength of a relation between terms is high (low) in data on the Web but the perceived strength of the relation is low (high), is the relation unexpected? To answer the question, we analyze unexpectedness using the data obtained by the method discussed in Section 6.4.2.

Figure 6.2 shows the results for all 770 pairs. Each dot corresponds to an attribute and object pair. Pairs with unexpectedness of < 3 are represented by various shades of blue (the darker the shade, the lower the unexpectedness). Pairs with unexpectedness of ≥ 3 are represented by various shades of red (the darker the shade, the higher the unexpectedness). The horizontal axis represents the normalized co-occurrence frequency on the Web. The vertical axis indicates the normalized perceived strengths of the relations between attributes and objects, which were obtained in Section 6.4.2.

In Figure 6.2, pairs with high unexpectedness occur in the lower right and upper left portions. The lower right (upper left) portion implies that the strength of the relation computed using the

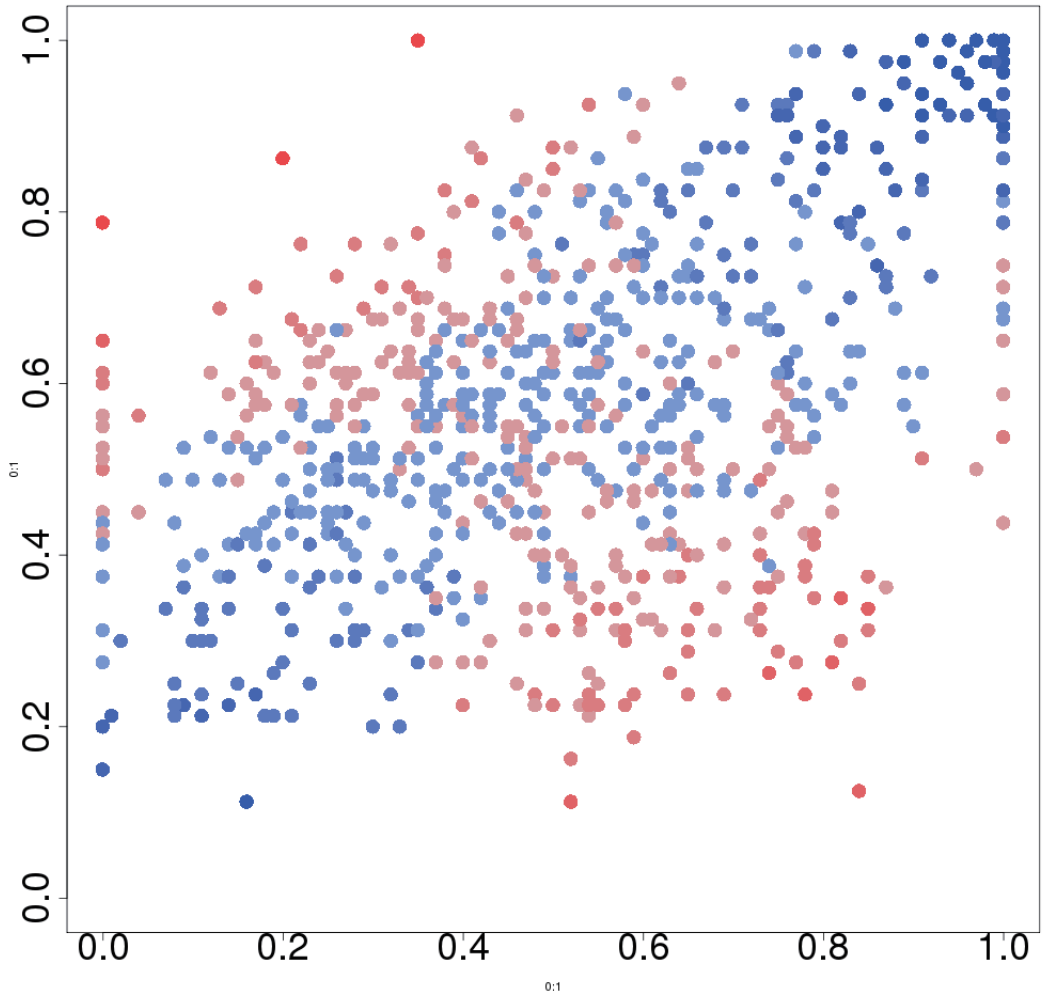

Figure 6.2: Distribution of degree of unexpectedness for all pairs (horizontal axis is normalized co-occurrence frequency; vertical axis is normalized perceived strength of the relation between an attribute and an object).

information on the Web is high (low) and the perceived strength of the relation is low (high). Pairs with low unexpectedness occur in the upper right and lower left portions. From these results, we deduce that when the gap between the strength of the relation computed using information on the Web and the perceived strength of the relation is very large, people feel that the information is unexpected.

Tables 6.3, 6.4, 6.5, and 6.6 show example data from the upper left, lower right, upper right, and lower left portions of Figure 6.2, respectively. For example, as seen in Table 6.3, many assessors answered that the relation between dietary fiber and a turnip was strong; however, the co-occurrence frequency on the Web was low because turnips are not rich in dietary fiber. Thus,

Table 6.3: Example data in the upper left portion.

attribute	object	perceived strength of the relation (1-5)	strength of the relation based on the co-occurrence frequency (0-100)	unexpectedness (1-5)
coffee	Argentina	3.65	22	3.8
Golden Glove Award	Tetsuharu Kawakami	3.5	31	3.3
dietary fiber	turnip	3.85	34	3.55
personal computer	BUFFALO	4.45	20	4.55

Table 6.4: Example data in the lower right portion.

attribute	object	perceived strength of the relation (1-5)	strength of the relation based on the co-occurrence frequency (0-100)	unexpectedness (1-5)
iron	chili pepper	2.5	64 3.55	
beer	China	2.7	73	3.45
mobile phone	YAMAHA	2.2	82	3.85
stolen base crown	Sadaharu Oh	2.05	74	4.05

Table 6.5: Example data in the upper right portion.

attribute	object	perceived strength of the relation (1-5)	strength of the relation based on the co-occurrence frequency (0-100)	unexpectedness (1-5)
pizza	America	4.7	76	2.0
protein	soybean	4.75	100	1.55
Kinkaku-ji Temple	Yoshimitsu Ashikaga	4.6	100	1.4
digital camera	EPSON	4.4	87	1.65

Table 6.6: Example data in the lower left portion.

attribute	object	perceived strength of the relation (1-5)	strength of the relation based on the co-occurrence frequency (0-100)	unexpectedness (1-5)
wine	Philippines	1.9	9	1.85
homo run king	Hiromichi Ishige	2.2	10	2.2
banana	Sweden	1.85	8	2.1
protein	eggplant	2.25	21	2.45

the assessors determined that the pair was unexpected due to this gap. Although a personal computer and BUFFALO have a strong relation, the co-occurrence frequency on the Web was low, which resulted in high unexpectedness. One way to solve this problem is to compute the strength of a relation between terms by considering context. This will enable us to discover more convincing unexpected information.

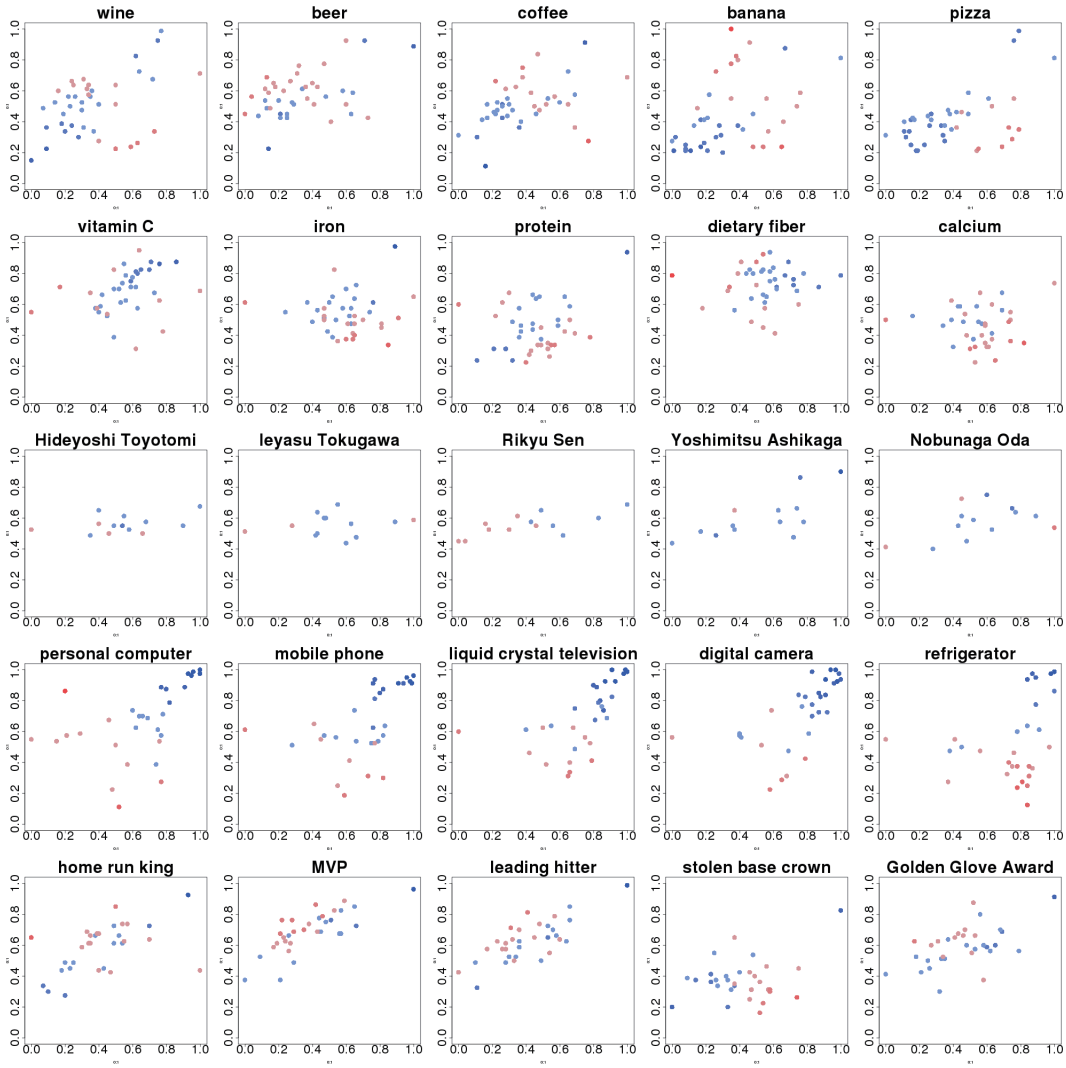

Figure 6.3: Distribution of degree of unexpectedness for each category (horizontal axis is normalized co-occurrence frequency; vertical axis is normalized perceived strength of relation between an attribute and an object).

Figure 6.3 shows the results for each attribute. For the attributes “pizza,” “banana,” “coffee,” “protein,” “refrigerator,” and “stolen base crown,” pairs with high unexpectedness are distributed in the lower right and upper left portions, which supports our hypotheses. Fn the category “tourist spot in Kyoto,” pairs with high unexpectedness are not distributed in the lower right or upper left portions because most assessors did not know whether each tourist spot has a strong relation with historical characters and most tourist spots are distributed at approximately 3.0 in terms of perceived strength.

6.4.4 Evaluation of Perceived Strength of a Relation

In this section, we answer research question 2: is considering the popularity of an object and similar objects of the object important to compute the perceived strength of the relation between the object and an attribute? We first describe the methods and evaluation metric used in this experiment and then report the results.

Methods and Evaluation Metric

We used the following methods to compute the perceived strength of a relation.

- Three existing methods introduced in Section 6.3.1, which are denoted by WLM, WebPMI, and Noda.
- Methods introduced in Section 6.3.2 that consider popularity of an object, which are denoted by WLM+pop, WebPMI+pop, and Noda+pop.
- Methods introduced in Section 6.3.3 that consider similar objects, which are denoted by WLM+BPR, WebPMI+BPR, and Noda+BPR.
- Methods introduced in Section 6.3.4 that consider the popularity of an object and similar objects, which are denoted by WLM+pop+BPR, WebPMI+pop+BPR, and Noda+pop+BPR.

Note that the damping factor ranges from 0.1 to 0.9 in increments of 0.1 in Equations 6.8 and 6.9.

To evaluate the methods for category and attribute pairs, we created two lists. The first is a list of objects in a category that are ranked in descending order of perceived strength of the relation with the attribute computed by each method. The second is a list of objects ranked in descending order of perceived strength of the relation with the attribute from crowdsourcing results. We then computed Spearman's rank correlation coefficient using these two lists.

Results

Figure 6.4 shows the average values of the correlation coefficient for all categories and those in each category. The y-axis represents the correlation coefficient. Here, "original" denotes WLM, WebPMI, or Noda. Damping factors that resulted in the highest correlation coefficient for each method are shown on each bar. With regard to the average values of all the categories, both pop and BPR outperformed all the existing methods. These results prove the effectiveness of Hypotheses 1 and 2 presented in Section 6.2. The pop+BPR method outperformed pop and BPR in all existing methods. These results indicate the effectiveness of considering both popularity

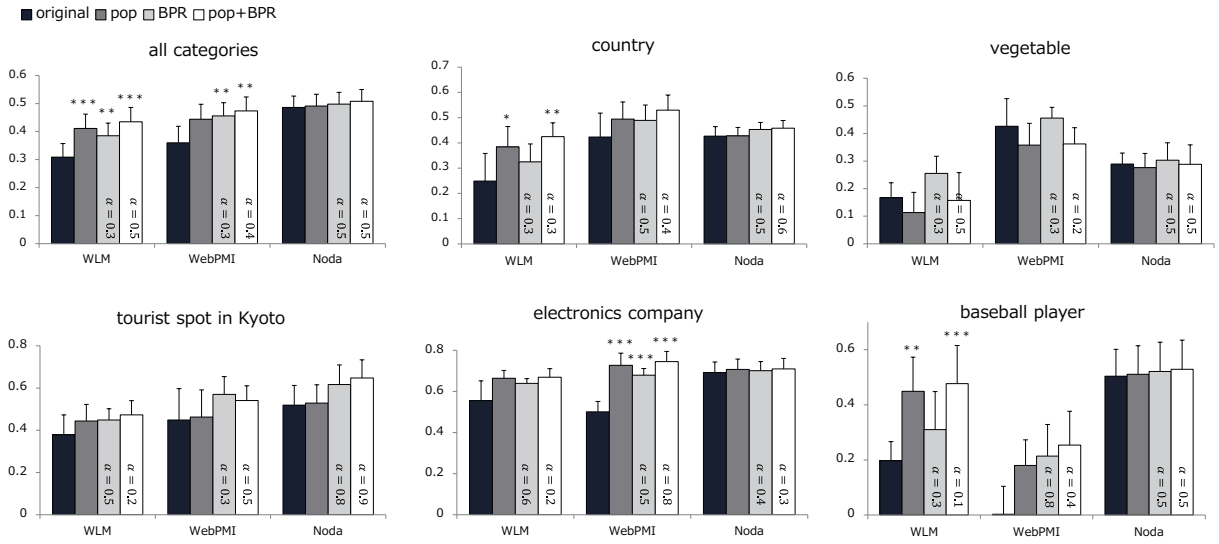

Figure 6.4: Average Spearman's rank correlation coefficient values for all methods in all categories and each category (α denotes damping factor; significant differences between the proposed methods and existing methods are denoted by $*$ ($\alpha = 0.1$), $**$ ($\alpha = 0.05$), and $***$ ($\alpha = 0.01$)).

of an object and similar objects. With the exception of the "vegetable" category, considering those two factors resulted in better correlation coefficients than the existing methods. For the "vegetable" category, Hypothesis 1 was invalid because the vegetables used in our experiments were very popular; popularity of an object had small impact on the results.

Figure 6.5 shows the average values of the correlation coefficient when the damping factor ranged from 0.1 to 0.9 in increments of 0.1 for methods that use the biased PageRank algorithm. In most categories, the correlation coefficient is reduced when the damping factor is very high, i.e., if the similarity of objects is given too much consideration, the accuracy of the results decreases.

Table 6.7 shows an example result (category is "country" and the attribute is "wine") for which considering similar objects had a positive impact. In this example the methods are Noda and Noda+BPR ($\alpha = 0.9$). The top 10 countries in terms of perceived strength of relation with wine are listed for each method. The numbers in parentheses show the rank from crowd sourcing results. In the Noda method, countries such as China, South Korea, and Thailand are listed in the top 10. However, most assessors answered that these countries are unrelated to wine, which indicates that the perceived strength of the relations between these countries and wine is low. Note that Taiwan and North Korea are appropriate coordinate terms of China, and the strength of the relation between these countries and wine computed by the Noda method is low. Therefore, the Noda+BPR method was able to move China down to 20th. Similarly, South Korea and Thailand

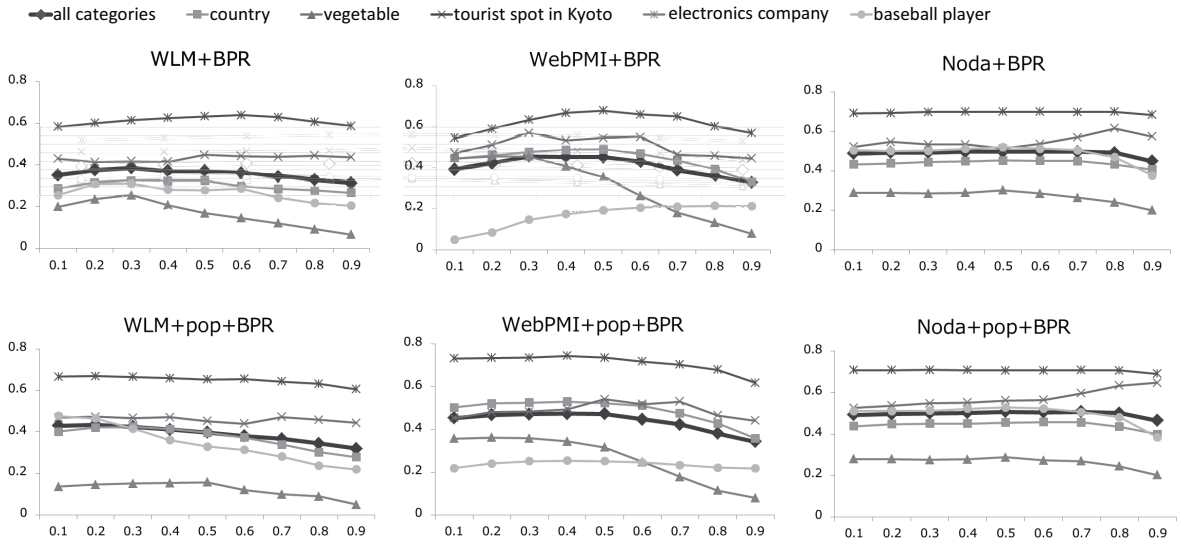

Figure 6.5: Average Spearman's rank correlation coefficient values when damping factor ranged from 0.1 to 0.9 in increments of 0.1 for methods that use the biased PageRank algorithm.

Table 6.7: Comparison of results from Noda method with the proposed method for category "country" and attribute "wine."

Rank	Noda	Noda+BPR $(\alpha = 0.9)$
1	France (1)	France (1)
2	Italy (2)	Italy (2)
3	Japan (5)	Japan (5)
4	Spain (3)	Spain (3)
5	Germany (4)	Germany (4)
6	the United States (6)	the United States (6)
7	China (32)	United Kingdom (11)
8	South Korea (36)	Australia (21)
9	Thailand (37)	Portugal (10)
10	Australia (21)	India (38)
correlation coefficient	0.466	0.702

moved down to 26th and 29th, respectively, when the Noda+BPR method was used.

Table 6.8 shows an example result (category is "electronics company" and the attribute is "liquid crystal television") for which considering both popularity of an object and similar objects had a positive effect. In this example, the methods are WebPMI and WebPMI+pop+BPR ($\alpha = 0.4$). The top 10 companies in terms of the perceived strength of the relation with liquid crystal televisions are listed for each method. The correlation coefficient was improved significantly from 0.442 to 0.838. Although "Seiko Epson" and "Victor Company of Japan" are ranked high

Table 6.8: Comparison of results from WebPMI method with the proposed method for category "electronics company" and attribute "liquid crystal television."

Rank	WebPMI	WebPMI+pop+BPR ($\alpha = 0.4$)
1	Sharp Corp. (2)	Sharp Corp. (2)
2	Seiko Epson (15)	Toshiba (3)
3	Victor Company of Japan (14)	Panasonic (1)
4	BUFFALO INC. (19)	Mitsubishi Electric (8)
5	Panasonic (1)	Sony (4)
6	Toshiba (3)	Hitachi,Ltd. (5)
7	KENWOOD (24)	Sanyo Electric (7)
8	Mitsubishi Electric (8)	Victor Company of Japan (14)
9	Sanyo Electric (7)	Fujitsu (6)
10	Daikin Industries,Ltd (30)	Nippon Electric Company (9)
correlation coefficient	0.442	0.838

in the list for the WebPMI method, the strength of the relation between these companies and liquid crystal televisions should be low because they do not produce liquid crystal televisions. One solution for this problem is to use methods that can more accurately compute the strength of a relation between terms, such as those proposed by Bollegala *et al.* [6] and Gabrilovich *et al.* [24].

6.4.5 Estimation of Unexpectedness of a Relation between Terms

In this section, we answer research question 3: given a pair of terms (object and attribute), can we estimate its unexpectedness? Here, we estimate unexpectedness on the basis of the popularity of an object t_{o_i} ($f_{pop}(t_{o_i})$), the perceived strength of the relation between an attribute a and o ($f_{rel}(t_a, t_{o_i})$), and the strength of the relation between a and o computed by the co-occurrence frequency ($f_{freq}(t_a, t_{o_i})$). Although several methods can estimate unexpectedness using these values, we used support vector regression (SVR) with the radial basis kernel function (RBF), which is the regression version of an SVM and is also used in some research in this field [47, 81]. Here, the objective variable is the unexpectedness of the a and o pair, and the explanatory variables are $f_{pop}(t_{o_i})$, $f_{cog}(t_a, t_{o_i})$, and $f_{freq}(t_a, t_{o_i})$. We used the SVR library LIBSVM [12]. To estimate unexpectedness, 25-fold cross validation over the 25 attributes was was performed. In each cross validation, we first learned parameters with 24 attributes, and then estimated the unexpectedness of each pair of the unused attribute and an object pair. We then computed the correlation coefficient between the estimated unexpectedness and the unexpectedness obtained by the method discussed in Section 6.4.2.

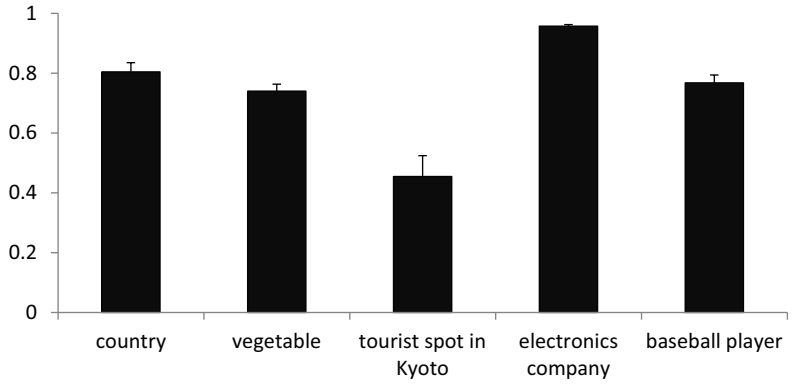

Figure 6.6: Correlation coefficient between unexpectedness judged by assessors and that estimated using only crowdsourcing results for perceived strength of relations.

The following three evaluations were conducted according to the type of $f_{rel}(t_a, t_{o_i})$ used in the training data and test data.

(1) Perceived strength of relations obtained by crowdsourcing was used for both training data and test data.

(2) Perceived strength of relations obtained by crowdsourcing was used for training data, and that computed by our proposed methods was used for test data.

(3) Perceived strength of relations computed by our proposed methods was used for both training data and test data.

For (1), we verify estimation accuracy of unexpectedness under an ideal situation in which the perceived strength of relations can be coumputed with 100% accuracy. For (2), we discuss the robustness of the perceived strength of relations obtained by crowdsourcing and the usefulness of estimating the perceived strength of relations with high accuracy. For (3), we do not use crowdsourcing results to verify whether we can estimate unexpectedness even when we do not obtain the perceived strength of relations using crowdsourcing.

Estimation of Unexpectedness Using Crowdsourcing Results

First, we estimated the unexpectedness for the attribute t_a and object t_{o_i} pair using only crowdsourcing results for $f_{rel}(t_a, t_{o_i})$. The results are shown in Figure 6.6. With the exception of the "tourist spot in Kyoto" category, correlation coefficients were high for all categories. They were particularly high for the "electronics company" category (up to 0.975). This is because the variance of the perceived strength of relations was high. These results indicate that unexpectedness

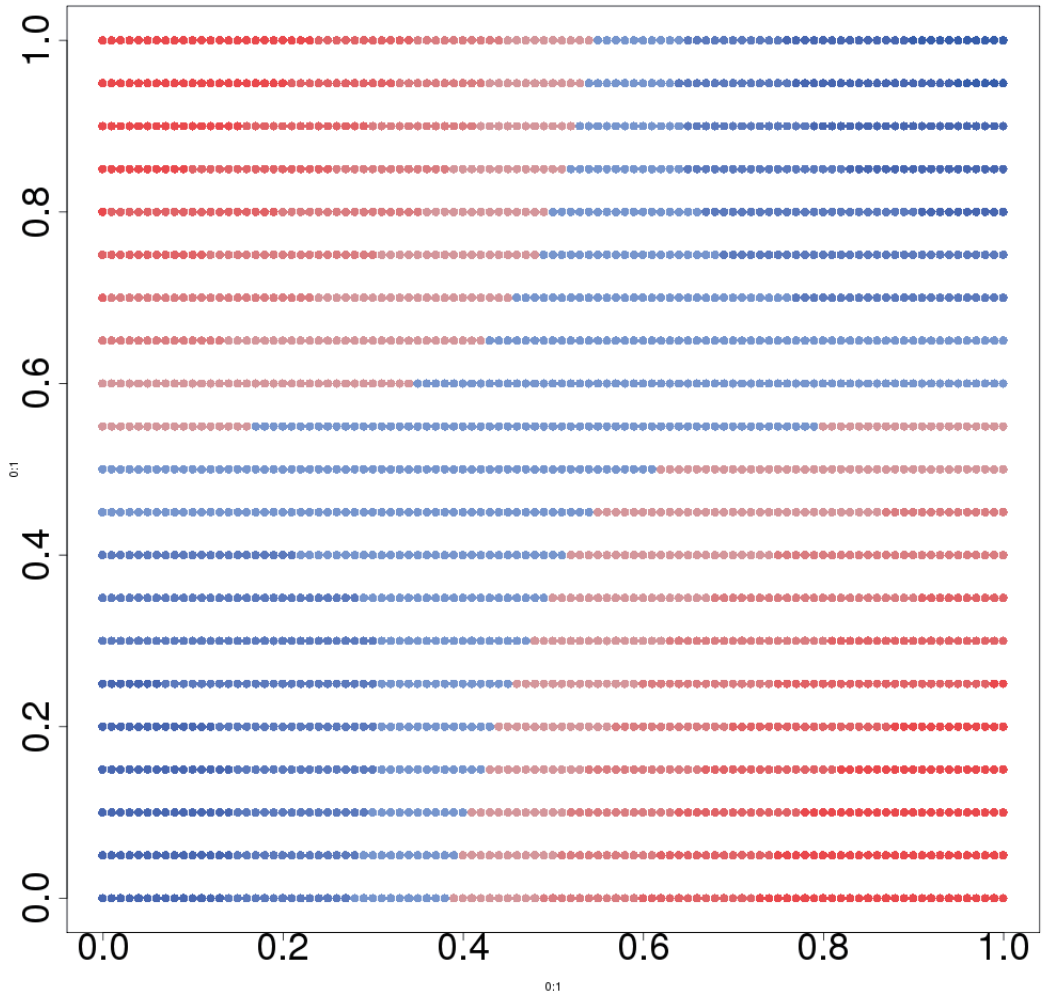

Figure 6.7: Distribution of unexpectedness estimated by SVR (horizontal axis is normalized co-occurrence frequency; vertical axis is normalized perceived strength of the relation between an attribute and object).

can be estimated with high accuracy when we can estimate the perceived strength of relations with high accuracy. For the "tourist spot in Kyoto" category, the correlation coefficient was moderate (0.455). One reason for this result is that the distribution of the perceived strength of relations in the category differed from other categories as indicated in Figure 6.3.

In addition to the above evaluation, we verified the distribution of unexpectedness. To achieve this objective, parameters were trained using all 770 pairs. Using the parameters, $f_{rel}(t_a, t_{o_i})$ ranged from 0 to 20 in increments of 1, and $f_{freq}(t_a, t_{o_i})$ ranged from 0 to 100 in increments of 1. The average popularity of all the objects was used for $f_{pop}(t_{o_i})$, . The unexpectedness for each $f_{rel}(t_a, t_{o_i})$ and $f_{freq}(t_a, t_{o_i})$ pair was also computed.

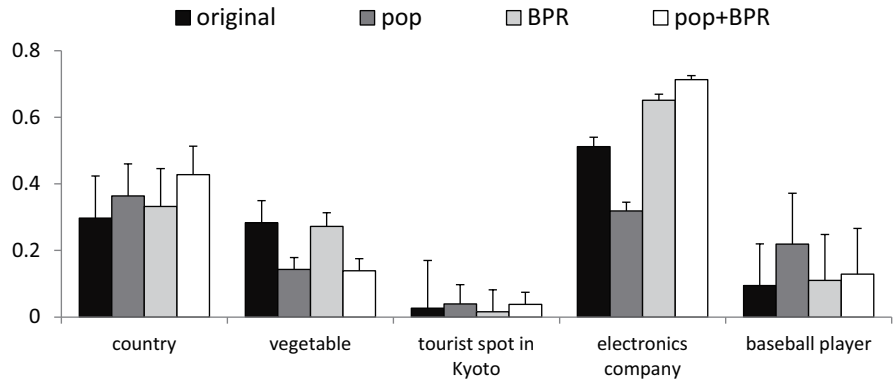

Figure 6.8: Correlation coefficient between unexpectedness judged by assessors and that estimated using crowdsourcing results and the proposed methods for perceived strength of relations.

The results are shown in Figure 6.7; dots are colored as in Figure 6.2. As can be seen, unexpectedness increases toward the upper left and lower right portions and decreases toward the upper right and lower left portions. From these results, we deduce that unexpectedness increases when the gap between the perceived strength of a relation and the strength in data on the Web is large.

Estimation of Unexpectedness Using Crowdsourcing Results and Proposed Methods

Next, we estimated the unexpectedness for an attribute t_a and object t_{o_i} pair using crowdsourcing results and the proposed methods for $f_{rel}(t_a, t_{o_i})$. As previously mentioned, crowdsourcing results were used for training data and our proposed methods were used for test data. The results are shown in Figure 6.8. For the "country" and "electronics company" categories, for which our proposed method estimated the perceived strength of relations with high accuracy, unexpectedness was also estimated with high accuracy. In these categories, estimation accuracy increased by considering the popularity of an object and similar objects. For the "vegetable" category, estimation accuracy of the perceived strength of relations decreased by considering the object's popularity in the experiment discussed in Section 6.4.4. It was observed that estimation accuracy of unexpectedness decreased by considering the object's popularity in the category. These results indicate that it is useful to compute the perceived strength of relations with high accuracy when estimating unexpectedness. For the "baseball player" category, the correlation coefficient was low because the estimation accuracy of the perceived strength of relations was low in Section 6.4.4.

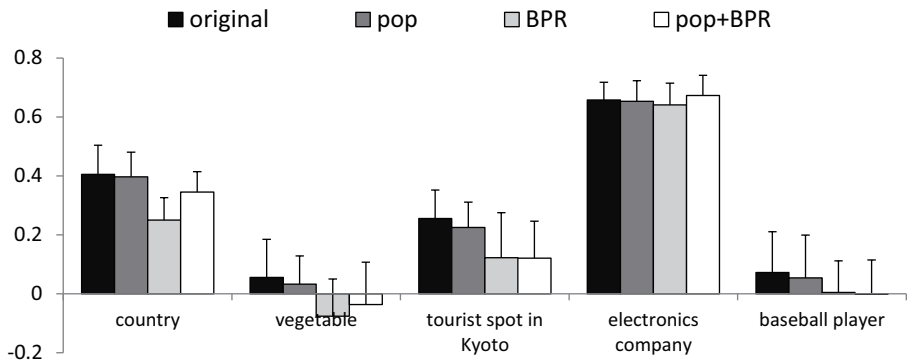

Figure 6.9: Correlation coefficient between unexpectedness judged by assessors and that estimated using only the proposed methods for perceived strength of relations.

Estimation of Unexpectedness Using Proposed Methods

Next, we estimated the unexpectedness for an attribute t_a and object t_{o_i} pair using only the proposed methods for $f_{rel}(t_a, t_{o_i})$. The results are shown in Figure 6.9. In this experiment, too, unexpectedness was estimated with high accuracy for the "country" and "electronics company" categories, for which our proposed method estimated the perceived strength of relations with high accuracy. However, we were unable to clarify the effectiveness of considering the popularity of an object and similar objects. When we used the perceived strength of relations computed by our proposed method in the training data, the computed values were inaccurate for some relations. Therefore, a useful model was not created by learning.

6.5 Summary

In this chapter, we focused on the strength of a relation computed by information on the Web and the perceived strength of the relation. We hypothesized that when the strength of a relation between terms is high (low) for data on the Web but the perceived strength of the relation is low (high), the relation is unexpected. To verify this hypothesis, we proposed a method for computing the perceived strength of a relation between terms (attribute and object). The proposed method considered two factors: (1) popularity of an object and (2) the strength of relations between an attribute and an object's coordinate terms. We conducted experiments using 25 attributes in five categories: country, vegetable, tourist spot in Kyoto, electronics company, and baseball player. We used crowdsourcing to collect data with regard to the perceived strength of the relation between an attribute and an object in order to evaluate the proposed method. The results showed the effectiveness of considering the aforementioned factors.

Our other experimental results also indicated that assessors perceived unexpectedness when

they knew there was a large gap between the perceived strength of the relation and the strength of the relation computed by data on the Web. We estimated the unexpectedness of a relation between terms on the basis of the popularity of an object, the perceived strength of the relation, and the strength of the relation computed by their co-occurrence frequency. The category "electronics company" achieved the highest correlation coefficient (0.792) between the human-judged unexpectedness and that estimated using SVR.

CHAPTER 7

CONCLUSIONS

7.1 Summary

This thesis discussed information retrieval techniques based on typicality and unexpectedness. We proposed methods for computing the typicality of an object set and the degree of unexpectedness of a relationship between terms. Four research topics addressed in this thesis are summarized as follows:

- **Search for an Object Set based on Typicality**
  We proposed a method for calculating the typicality of an object set such as a recipe and a tourist route. An object set consists of some objects such as ingredients and tourist spots. The proposed method first detected the most typical set of objects in a category based on the appearance frequency of each object and the co-occurrence frequencies between objects. Given an object set, we computed the degree of its typicality based on the affinity between its objects and the difference between the object set and the most typical set of objects. We also proposed methods for recommending candidate objects for addition to and deletion from an object set to change it to a more typical or atypical set. We focused on recipes as object sets and conducted experiments. The results showed that the correlation coefficient between human-judged typicality and that computed by the proposed method was as high as 0.868 in a category. In the experiment regarding the addition and deletion of ingredients, we found that the proposed method was particularly effective in recommending the addition and deletion of ingredients to change a recipe to a more atypical one. We also focused on two characteristics of typicality that have been proposed in cognitive psychology. We used recipe data and compared the degree of typicality of a recipe judged by assessors with that calculated based on each characteristic. Evaluation experiments showed that a

characteristic based on similarity was able to estimate the typicality judged by assessors with high accuracy in a category in which the similarity of properties between objects was high.

- **Ranking of Coordinate Terms and Hypernyms Using a Hypernym-Hyponym Dictionary**

  We proposed methods for ranking coordinate terms and hypernyms of a given query according to their appropriateness. In the proposed method, a bipartite graph was created based on hypernyms of a query and hyponyms of each hypernym using a hypernym-hyponym dictionary. Subsequently, we applied a HITS-based algorithm to the bipartite graph and ranked coordinate terms and hypernyms based on their appropriateness. The experimental results obtained using 50 queries demonstrated that our method could rank appropriate coordinate terms and hypernyms higher than other comparable methods.

- **Discovering Unexpected Information based on the Popularity of Terms and the Typicality of Relationships between Terms**

  We proposed a method for discovering unexpected information for a given query. Given a query q (e.g., "Hiromitsu Ochiai"), our method first detected an unexpected related term e (e.g., "Gundam") and then presented unexpected information (e.g., "Hiromitsu Ochiai is a Gundam maniac."). We hypothesized that information was unexpected when it included a related term that had an atypical relationship with the query and the degree of popularity of the related term is high. We compute the typicality of the relationship between a query and its related term based on the relationships between their coordinate terms using Wikipedia data. We conducted an experiment using 75 queries in five domains, i.e., the names of people, regions, products, facilities, and organizations. The 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.

- **Discovering an Unexpected Relationship by Measuring Perceived Strength of the Relationship between Terms**

  We focused on the difference between the strength of the relationship between terms for information receivers and that for information senders. We hypothesized that when the strength of the relationship between terms is high (low) for information receivers but low (high) for information senders, the information may be unexpected. To verify this hypothesis, we proposed a method for computing perceived strength of the relationship between

terms (an attribute and an object) for information receivers. The proposed method considered two factors: (1) the popularity of an object, and (2) the strength of the relationships between an attribute and an object's coordinate terms. We conducted experiments using 25 attributes that were included in five categories, i.e., country, vegetable, tourist spot in Kyoto, electronic company, and baseball player. We utilized crowd sourcing to collect data for the perceived strength of the relationship between an attribute and an object and evaluated the proposed method. The results showed the effectiveness of considering the popularity of an object and the strength of the relationships between attributes and coordinate terms. Our experimental results also indicated that assessors considered information to be unexpected when they knew that there was a gap between the strength of the relationship for information receivers and that for information senders. We estimated the unexpectedness of the relationship between terms based on the popularity of an object, the perceived strength of the relationship for information receivers, and the strength of the relationship for information senders. The category "electronic company" achieved the highest correlation coefficient of 0.792 between human-judged degree of unexpectedness and that computed by our method.

7.2 Future Directions

There are several research topics that need to be explored in the future. First, we would like to consider the diversity of information sender in an information source. For example, assume that there are many similar objects to an object, but the object's social recognition degree is low. If the characteristics of information senders are similar, the diversity of information senders is low. In such a case, it is not surprising that the object's social recognition degree is low because only people in a certain community send the information. Conversely, if characteristics of information senders differ, the diversity of information senders is high. In such a case, it is surprising, or unexpected, that the object's social recognition degree is low even though information is sent by people in various communities. Second, we would like to consider the other concepts of typicality. In this research, of the three typicality concepts that were discussed in Barsalou's study [4], we focused on central tendency and the frequency of instantiation. By considering the third concept, or ideals, we are able to discover greater variety of unexpected information. For example, assume an object is very similar to a goal associated with its category and is typical from the viewpoint of ideals. If the social recognition degree of the object is low and the object is atypical from the viewpoint of the frequency of instantiation, the object is an unexpected object. Conversely, if an object is typical due to a high degree of social recognition but atypical from the viewpoint of ideals, the object is also an unexpected object. Considering these factors enables us

to discover and utilize various types of atypical and useful information.

BIBLIOGRAPHY

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pp. 5–14, 2009.

[2] K. Akamatsu, N. Pattanasri, A. Jatowt, and K. Tanaka. Measuring comprehensibility of web pages based on link analysis. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '11, pp. 40–46, 2011.

[3] L. Barsalou. Ad hoc categories. *Memory & cognition*, 11(3):211–227, 1983.

[4] L. Barsalou. Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, pp. 629–654, 1985.

[5] 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*, pp. 281–309, 1998.

[6] D. Bollegala, Y. Matsuo, and M. Ishizuka. A web search engine-based approach to measure semantic similarity between words. *IEEE Trans. on Knowl. and Data Eng.*, 23(7):977–990, 2011.

[7] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International Conference on World Wide Web*, WWW '98, pp. 107–117, 1998.

[8] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[9] Y. Cai and H.-f. Leung. Formalizing object typicality in context-aware ontology. In *Proceedings of the 2008 20th IEEE International Conference on Tools with Artificial Intelligence - Volume 02*, ICTAI '08, pp. 233–240, 2008.

[10] Y. Cai, H.-f. Leung, Q. Li, J. Tang, and J. Li. Recommendation based on object typicality. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pp. 1529–1532, 2010.

[11] P. Chandar and B. Carterette. Using preference judgments for novel document retrieval. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pp. 861–870, 2012.

[12] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[13] H.-H. Chen, M.-S. Lin, and Y.-C. Wei. Novel association measures using web search with double checking. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pp. 1009–1016, 2006.

[14] S. Cheng, A. Arvanitis, and V. Hristidis. How fresh do you want your search results? In *Proceedings of the 22Nd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '13, pp. 1271–1280, 2013.

[15] J. Cho and S. Roy. Impact of search engines on page popularity. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pp. 20–29, 2004.

[16] J. Cho, S. Roy, and R. E. Adams. Page quality: In search of an unbiased web ranking. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, SIGMOD '05, pp. 551–562, 2005.

[17] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma. Probabilistic query expansion using query logs. In *Proceedings of the 11th International Conference on World Wide Web*, WWW '02, pp. 325–332, 2002.

[18] H. Deng, M. R. Lyu, and I. King. A generalized Co-HITS algorithm and its application to bipartite graphs. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pp. 239–248, 2009.

[19] A. Dong, Y. Chang, Z. Zheng, G. Mishne, J. Bai, R. Zhang, K. Buchner, C. Liao, and F. Diaz. Towards recency ranking in web search. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pp. 11–20, 2010.

[20] Z. Dou, S. Hu, K. Chen, R. Song, and J.-R. Wen. Multi-dimensional search result diversification. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pp. 475–484, 2011.

[21] C. Fellbaum. *WordNet: An Electronic Lexical Database*. Mit Press, 1998.

[22] J. L. Fleiss and J. Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33:613–619, 1973.

[23] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Trans. Inf. Syst.*, 23(2):147–168, 2005.

[24] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pp. 1606–1611, 2007.

[25] M. Ge, C. Delgado-Battenfeld, and D. Jannach. Beyond accuracy: Evaluating recommender systems by coverage and serendipity. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, pp. 257–260, 2010.

[26] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, VLDB '04, pp. 576–587, 2004.

[27] A. Hassan and R. W. White. Personalized models of search satisfaction. In *Proceedings of the 22nd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '13, 2013.

[28] T. H. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the 11th International Conference on World Wide Web*, WWW '02, pp. 517–526, 2002.

[29] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, pp. 539–545, 1992.

[30] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, 2004.

[31] L. Iaquinta, M. de Gemmis, P. Lops, G. Semeraro, M. Filannino, and P. Molino. Introducing serendipity in a content-based recommender system. In *Proceedings of the 2008 8th International Conference on Hybrid Intelligent Systems*, HIS '08, pp. 168–173, 2008.

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

[33] Y. Jing and S. Baluja. Visualrank: Applying pagerank to large-scale image search. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1877–1890, 2008.

[34] M. P. Kato, T. Yamamoto, H. Ohshima, and K. Tanaka. Investigating users' query formulations for cognitive search intents. In *Proceedings of of the 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, 2014.

[35] H. Kawai, H. Mizuguchi, and M. Tsuchida. Cost-effective web search in bootstrapping for named entity recognition. In *Database Systems for Advanced Applications*, Vol. 4947 of *Lecture Notes in Computer Science*, pp. 393–407, 2008.

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

[37] R. Lempel and S. Moran. SALSA: the stochastic approach for link-structure analysis. *ACM Trans. Inf. Syst.*, 19:131–160, 2001.

[38] M.-J. Lesot, L. Mouillet, and B. Bouchon-Meunier. Fuzzy prototypes based on typicality degrees. In *Computational Intelligence, Theory and Applications*, Vol. 33 of *Advances in Soft Computing*, pp. 125–138. 2005.

[39] B. Liu, Y. Ma, and P. S. Yu. Discovering unexpected information from your competitors' web sites. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, pp. 144–153, 2001.

[40] G. Lu, P. Huang, L. He, C. Cu, and X. Li. A new semantic similarity measuring method based on web search engines. *W. Trans. on Comp.*, 9(1):1–10, 2010.

[41] D. Medin and E. Smith. Concepts and concept formation. *Annual Review of Psychology*, 35:113–138, 1984.

[42] Y. Mejova, I. Bordino, M. Lalmas, and A. Gionis. Searching for interestingness in wikipedia and yahoo!: answers. In *Proceedings of the 22nd International Conference on World Wide Web Companion*, WWW '13 Companion, pp. 145–146, 2013.

[43] C. B. Mervis and J. R. Pani. Acquisition of basic object categories. *Cognitive Psychology*, 12(4):496–522, 1980.

[44] R. Mihalcea and P. Tarau. Textrank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Vol. 4 of *EMNLP '04*, pp. 404–411, 2004.

[45] G. A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[46] D. Milne and I. Witten. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceedings of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, pp. 25–30, 2008.

[47] S. Moghaddam, M. Jamali, and M. Ester. Review recommendation: Personalized prediction of the quality of online reviews. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pp. 2249–2252, 2011.

[48] G. L. Murphy. The big book of concepts. *MIT Press*, 2002.

[49] A. Nadamoto, E. Aramaki, T. Abekawa, and Y. Murakami. Searching for important but neglected content from community-type-content. In *Proceedings of the Fourth International Conference On Signal-Image Technology and Internet-based Systems*, SITIS '08, pp. 161–168, 2008.

[50] A. Nadamoto, E. Aramaki, T. Abekawa, and Y. Murakami. Content hole search in community-type content. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pp. 1223–1224, 2009.

[51] S. Nakamura, S. Konishi, A. Jatowt, H. Ohshima, H. Kondo, T. Tezuka, S. Oyama, and K. Tanaka. Trustworthiness analysis of web search results. In *Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL'07, pp. 38–49, 2007.

[52] M. Nakatani, A. Jatowt, and K. Tanaka. Easiest-first search: Towards comprehension-based web search. In *Proceedings of the 18th ACM International Conference on Information and Knowledge Management*, CIKM '09, pp. 2057–2060, 2009.

[53] T. Noda, H. Ohshima, S. Oyama, K. Tajima, and K. Tanaka. 主題語からの話題語自動抽出とこれに基づく web 情報検索. *DBSJ Letters*, 5(2):69–72, 2006 (in Japanese).

[54] Y. Noda, Y. Kiyota, and H. Nakagawa. Discovering serendipitous information from wikipedia by using its network structure. In *Proceedings of Fourth International AAAI Conference on Weblogs and Social Media*, ICWSM '10, pp. 299–302, 2010.

[55] H. Ohshima, S. Oyama, and K. Tanaka. Searching coordinate terms with their context from the web. In *Proceedings of the Seventh International Conference on Web Information Systems*, WISE'06, pp. 40–47, 2006.

[56] K. Oku and F. Hattori. Fusion-based recommender system for improving serendipity. In *Proceedings of the Workshop on Novelty and Diversity in Recommender Systems, at the Fifth ACM International Conference on Recommender Systems*, DiveRS '11, RecSys '11, pp. 19–26, 2011.

[57] B. Padmanabhan and A. Tuzhilin. A belief-driven method for discovering unexpected patterns. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, SIGKDD '98, pp. 94–100, 1998.

[58] B. Padmanabhan and A. Tuzhilin. Small is beautiful: discovering the minimal set of unexpected patterns. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '00, pp. 54–63, 2000.

[59] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pp. 275–281, 1998.

[60] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, CSCW '94, pp. 175–186, 1994.

[61] M. Rifqi. Constructing prototypes from large databases. In *IN IPMU'96*, pp. 301–306, 1996.

[62] A. Ritter, S. Soderland, and O. Etzioni. What is this, anyway: Automatic hypernym discovery. In *AAAI Spring Symposium: Learning by Reading and Learning to Read*, pp. 88–93, 2009.

[63] S. E. Robertson and K. S. Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146, 1976.

[64] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pp. 232–241, 1994.

[65] D. E. Rose and D. Levinson. Understanding user goals in web search. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pp. 13–19, 2004.

[66] M. Sahami and T. D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pp. 377–386, 2006.

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

[68] R. L. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 881–890, 2010.

[69] K. Shinzato and K. Torisawa. Acquiring hyponymy relations from web documents. In *HLT-NAACL*, pp. 73–80, 2004.

[70] R. Snow, D. Jurafsky, and A. Y. Ng. Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems 17*, pp. 1297–1304. 2005.

[71] M. Strube and S. P. Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, pp. 1419–1424, 2006.

[72] K. M. Svore and C. J. Burges. A machine learning approach for improved BM25 retrieval. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pp. 1811–1814, 2009.

[73] H. Taher, K. Sepandar, and J. Glen. An analytical comparison of approaches to personalizing pagerank. In *Stanford University Technical Report 2003*, 2003.

[74] S. Tanaka, A. Jatowt, M. P. Kato, and K. Tanaka. Estimating content concreteness for finding comprehensible documents. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pp. 475–484, 2013.

[75] A. Tuzhilin. On subjective measures of interestingness in knowledge discovery. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, SIGKDD '95, pp. 275–281, 1995.

[76] R. C. Wang and W. W. Cohen. Language-independent set expansion of named entities using the web. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, pp. 342–350, 2007.

[77] J.-R. Wen, J.-Y. Nie, and H.-J. Zhang. Clustering user queries of a search engine. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pp. 162–168, 2001.

[78] M. Yamaguchi, H. Ohshima, S. Oyama, and K. Tanaka. Unsupervised discovery of coordinate terms for multiple aspects from search engine query logs. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '08, pp. 249–257, 2008.

[79] C. Yeung and H. Leung. Ontology with likeliness and typicality of objects in concepts. *Conceptual Modeling-ER 2006*, pp. 98–111, 2006.

[80] H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pp. 129–136, 2003.

[81] Z. Zhang and B. Varadarajan. Utility scoring of product reviews. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, pp. 51–57, 2006.

[82] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pp. 22–32, 2005.

PUBLICATIONS

Journal Papers

(in Japanese)

1. 佃洸摂，大島裕明，山本光穂，岩崎弘利，田中克己
   “語の認知度と語間の関係の非典型度に基づく Wikipedia からの意外な情報の発見”
   情報処理学会論文誌:データベース *(TOD 61)*, Vol.7, No.1, pp.1–17, 2014.

2. 佃洸摂，大島裕明，加藤誠，田中克己
   “属性値の同一性・相補性に着目したオブジェクト集合検索手法の提案とその観光地データへの適用”
   情報処理学会論文誌:データベース *(TOD 60)*, Vol.6, No.5, pp.49–61, 2013.

3. 佃洸摂，大島裕明，山本光穂，岩崎弘利，田中克己
   “語の認知度と同位語間の関係に基づく意外な情報の発見”
   日本データベース学会論文誌 *(DBSJ Journal)*, Vol.11, No.3, pp.21–26, 2013.

4. 佃洸摂，中村聡史，山本岳洋，田中克己
   “映像に付与されたコメントを用いた登場人物が注目されるシーンの推定”
   情報処理学会論文誌「情報爆発」特集号, Vol.52, No.12, pp.3471–3482, 2011.

5. 佃洸摂，中村聡史，山本岳洋，田中克己
   “レシピ検索のためのレシピの構造とその安定度を考慮した追加・削除可能な食材の推薦”
   電子情報通信学会和文論文誌 *A* 料理を取り巻く情報メディア技術特集号, Vol. J94-A, No. 7, pp.476–487, 2011.

International Conference Papers

1. Kosetsu Tsukuda, Hiroaki Ohshima, Katsumi Tanaka
   “Ranking of Coordinate Terms and Hypernyms Using a Hypernym-Hyponym Dictionary”

In *Proceedings of the 2014 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2014)*, pp.15–21, 2014.

2. Kosetsu Tsukuda, Tetsuya Sakai, Zhicheng Dou, Katsumi Tanaka
   "Estimating Intent Types for Search Result Diversification"
   In *Proceedings of the 9th Asia Information Retrieval Societies Conference (AIRS 2013)*, pp.25–37, 2013.

3. Kosetsu Tsukuda, Hiroaki Ohshima, Mitsuo Yamamoto, Hirotoshi Iwasaki, Katsumi Tanaka
   "Discovering Unexpected Information on the basis of Popularity/Unpopularity Analysis of Coordinate Objects and their Relationships"
   In *Proceedings of the 28th ACM Symposium On Applied Computing (SAC 2013)*, pp. 878–885, 2013

4. Kosetsu Tsukuda, Takehiro Yamamoto, Satoshi Nakamura, Katsumi Tanaka
   "Plus One or Minus One: A Method to Browse from an Object to Another Object by Adding or Deleting an Element"
   In *Proceedings of the 21st International Conference on Database and Expert Systems Applications (DEXA2010)*, pp. 258–266, 2010

International Workshop Papers

1. Kosetsu Tsukuda, Zhicheng Dou, Tetsuya Sakai, Katsumi Tanaka
   "Microsoft Research Asia at the NTCIR-10 Intent Task"
   In *Proceedings of the 10th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-10)*, pp. 152–158, 2013

2. Tomohiro Manabe, Kosetsu Tsukuda, Kazutoshi Umemoto, Yoshiyuki Shoji, Makoto P. Kato, Takehiro Yamamoto, Meng Zhao, Soungwoong Yoon, Hiroaki Ohshima, Katsumi Tanaka
   "Information Extraction based Approach for the NTCIR-10 1CLICK-2 Task"
   In *Proceedings of the 10th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-10)*, pp. 243–249, 2013

3. Makoto P. Kato, Meng Zhao, Kosetsu Tsukuda, Yoshiyuki Shoji, Takehiro Yamamoto, Hiroaki Ohshima, Katsumi Tanaka
   "Information Extraction based Approach for the NTCIR-9 1CLICK Task"
   In *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-9)*, pp. 202–207, 2011

Domestic Symposium and Workshops

(in Japanese)

1. 佃洸摂，大島裕明，加藤誠，田中克己
   “オブジェクト間の意外な共通点の発見”
   第 6 回データ工学と情報マネジメントに関するフォーラム (DEIM2014), 2014

2. 佃洸摂，大島裕明，田中克己
   “上位下位概念辞書を用いた同位語・上位語のランキング手法の提案”
   Web とデータベースに関するフォーラム (WebDB Forum 2013), 2013

3. 佃洸摂，大島裕明，山本光穂，近藤賢志，田中克己
   “属性の組み合わせとその相性に基づく複合オブジェクト検索”
   第 2 回 Web インテリジェンスとインタラクション研究会 (第 2 回 WI2), 2013

4. 佃洸摂，大島裕明，山本光穂，近藤賢志，田中克己
   “属性のマッチングに基づくオブジェクトの相性検索”
   第 5 回データ工学と情報マネジメントに関するフォーラム (DEIM2013), 2013

5. 佃洸摂，大島裕明，山本光穂，岩崎弘利，田中克己
   “エンティティ間の関連性に基づく意外な情報の発見”
   第 4 回データ工学と情報マネジメントに関するフォーラム (DEIM2012), 2012

6. 佃洸摂，大島裕明，山本光穂，岩崎弘利，田中克己
   “語の認知度と同位語間の関係に基づく意外な情報の発見”
   Web とデータベースに関するフォーラム (WebDB Forum 2012), 2012

7. 佃洸摂，中村聡史，田中克己
   “視聴者の反応に基づく動画検索および推薦システムの提案”
   第 19 回インタラクティブシステムとソフトウェアに関するワークショップ (WISS2011), 2011

8. 佃洸摂，中村聡史，山本岳洋，田中克己
   “オブジェクトの典型度分析とその検索への応用”
   Web とデータベースに関するフォーラム (WebDB Forum 2011), 2011

9. 佃洸摂，中村聡史，田中克己
   “時刻同期コメントを用いた動画の登場人物の活躍度推定とその検索への応用”
   第 144 回ヒューマンコンピュータインタラクション研究会, 2011

10. 佃洸摂，中村聡史，山本岳洋，田中克己
“ソーシャルアノテーションに基づく動画の登場人物の重要度の推定”
Web とデータベースに関するフォーラム (WebDB Forum 2010), 2010

11. 佃洸摂，中村聡史，山本岳洋，田中克己
“動画インデックスのための登場人物の主役度および脇役度の推定”
第 18 回 Web インテリジェンスとインタラクション研究会, 2010

12. 佃洸摂，中村聡史，田中克己
“集合型 Web オブジェクトの構成要素に対する追加・削除要素の推薦とそのレシピデータへの応用”
第 2 回データ工学と情報マネジメントに関するフォーラム (DEIM2010), 2010

13. 佃洸摂，中村聡史，田中克己
“エンティティの構成要素の安定性を考慮したインタラクティブな Web エンティティ検索とそのレシピ推薦への応用”
情報処理学会創立 50 周年記念全国大会, 2010