A Study on Social Information Search and Analysis on the Web by Diversity Computation

Yoshiyuki Shoji
Graduate School of Informatics,
Kyoto University, Japan

A Study on Social Information Search and Analysis on the Web by Diversity Computation

Doctoral Dissertation Series of Tanaka Laboratory
Department of Social Informatics,
Graduate School of Informatics,
Kyoto University

Abstract

The usage of the World Wide Web (Web) becomes more diverse with the increase of the number of users. Nowadays, the Web has aspects of social media, *e.g.*, communication media and commercial media. Existing Web search systems rank search results on the basis of the term frequency in Web page contents and the number of hyperlinks from/to Web pages. However, these features are not sufficient to reflect the social aspects of the current Web. In this thesis, we propose three search methods which consider social backgrounds of Web pages, *i.e.*, who focus on pages, why they focus on pages and how diverse they are. This thesis discusses the Web information search and diversity analysis techniques on the social information including the following three research topics:

1. **Search-by-Reaction:**
   Web Search Using People's Reaction Terms in Twitter
   We describe a new concept for improving Web search performance using Web 1.0 and Web 2.0 contents in a complementary manner. In particular, we describe a social media search, called Search-by-Reaction, which uses posts on microblogs such as Twitter. Conventional Web search engines suffer from low precision and recall when a user query contains impressions such as "cute" or evaluations such as "easy-to-understand" as keywords. Web search users often wish to input such impression keyword terms together with ordinary topic keyword terms. The difficulty in Web page searches using these impression terms or evaluation terms comes from the fact that the target Web pages do not always contain them. For example, conventional Web search engines cannot necessarily find an "easy-to-understand document about C-language" given the query "easy-to-understand C-language" because they only focus on terms appearing on pages. A possible way to cope with this problem is to exploit user tag data attached to corresponding Web pages. There is, however, still a serious problem with this idea in that there is not enough annotation information of people's impressions and evaluations of the target Web pages. Our idea of search-

by-reaction focuses on the terms people use when they find pages they like, and we focus on Web communications on social networking site (SNS) such as Twitter. Our system extracts such reaction terms from Twitter by analyzing the impression terms or evaluation terms specified in user queries, and it uses the extracted reaction terms to search for Web pages.

2. **Diversity-Based Credibility Analysis:**
   Can Diversity Improve Credibility of User Review Data?
   In this research, we propose methods of estimating the credibility of reviewers as individuals and as groups, where the credibility is defined as the ability of precisely estimating the quality of items. Our proposed methods are built on two simple hypotheses: 1) a reviewer who has reviewed many and diverse items has high credibility, and 2) a group of reviewers is credible if the group consists of many and diverse reviewers. To verify the two assumptions, we conducted experiments with a movie review dataset. The experimental results showed that the diversity of reviewed items and reviewers was effective to estimate the credibility of reviewers and reviewer groups, respectively. Therefore, yes, the diversity does improve the credibility of user review data.

3. **Diversity-Based HITS:**
   Web Page Ranking by Referrer and Referral Diversity
   We propose a Web page ranking method that considers the diversity of linked pages and linking pages. Typical link analysis algorithms such as HITS and PageRank calculate scores of pages on the basis of the number of link between pages. However, even if the number of links is the same, there is often a big difference between documents linked by pages with similar contents and those linked by pages with different contents. We propose two types of link diversity, referral diversity (diversity of pages linked by the page) and referrer diversity (diversity of pages linking to the page)to calculate diversity scores of pages, aiming at expanding the basic HITS algorithm. The results of repeated experiments showed that the diversity-based method was more useful than the original HITS algorithm for finding suitable information on the Web for ordinary users.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Background

The World Wide Web (Web) is not a mere collection of documents, but it also functions as an essential part of our social life nowadays. With the rapid growth of the Internet, the usage of the Web is becoming more multi-faceted. For instance, the Web has been widely regarded as a communication tool, with people from all over the world using Social Networking Services (SNS) to communicate with each other. According to the result of a survey by the Ministry of Internal Affairs and Communications of Japan*, 41.3% of the Internet users use SNSs too. Thus, the Web is currently a place where a considerable amount of communication log data are recorded, aggregated and investigated. Furthermore, electronic commerce showed an increasing prosperity in recent years. The aforementioned survey also demonstrated that almost 56.9% of users have experiences of shopping online by using the Web. The shopping mode of e-commerce allows users to compare one product with similar ones and competitors through product descriptions and historical reviews from different resources. After purchasing, the user may also give feedback to the target product, which may influence other people's purchasing decision in the future. As illustrated above, the Web continues offering an increasing number of different online services in order to meet the requirements of distinct social needs, which contributes to the creation of diverse Web resources. In addition, the survey also showed that the number of users of the Web became nearly three billions in 2014. It appears to reflect that the current Web lowered the threshold of people accessing the services provided online, realizing the dream that not only

*http://www.soumu.go.jp/johotsusintokei/statistics/

expert users of computers, but also common people can utilize the Web. Correspondingly, the online services ought to overcome the challenge brought by the characteristic of the diversity of users. Generally, the importance of the Web search tends to focus on the way of diversifying the resources, the usage and the users.

However, existing Web search algorithms are not suitable enough for the current Web environment. Models in most of these algorithms (*e.g.*, vector space model and link analysis model) focus only on the contents of Web documents and the number of links between them, but not on their social background. A great number of the contents on the Web are generated by a huge number of diverse people, and documents are written and connected to each other for different purposes. Thus, all the documents and links on the Web have background information: the reason and aim of the documents and links themselves, the property of the users who create these links and contents, the structure of document set and so forth. In this thesis, we propose a methodology which focuses on the background of the documents and links.

1.2 Approach

To focus on the background information of Web pages and links, we proposed two types of approaches. The first approach is to use links between ordinary Web 1.0 sites and Web 2.0 sites. Existing Web search models did not distinguish links in the social media and links in the ordinary Web. In social media, there are some special links, *e.g.*, links between users' opinion and pages, and links between users and items. Our methods use these special links to take *how* information and *who* information as the background of links. Another approach is to analyze diversity of the Web users. To utilize information taken from social media, we must consider the social behavior of users. Our methods use diversity analysis to take *who* and *how* as the background information of links.

Existing search models cannot exploit the benefits of the Web 2.0 that are originated by the socialization of the Web. The recent Web environment contains richer and more diverse information which did not appear in the old static Web. A typical example is the opinion of readers. In the ordinary Web, users that want to submit their opinion had to overcome some difficult technical hurdles, *i.e.*, they must write static HTML document, they must make the server space up, and they must upload and structure the Web site. Under such circumstances, opinions of novices or minor opinions were very scarce on the Web. Nevertheless, in the current Web everyone can upload his opinions lightheartedly and easily. In social media services, *e.g.*, SNS sites like Twitter or Facebook, and CGM sites like review sites, we can find a massive amount of trivial opinions by non-expert users. Our

goal is to use these trivial information by common people in Web 2.0 as the information source of the Web search and the Web analysis.

The Search-by-Reaction, which is our first work, uses such kind of social data to improve the Web search. However, it brings out the weakness of social data. The major problem of using such kind of resources is the information quality, as these resources made by a lot of diverse common people. Some of them are novice users or just liars; and their opinions may be inaccurate. Another problem is that they create such resources through their social activities. These information is often influenced by social factors. One of the most typical examples of social factor is the properties of the groups. Some of groups have an orientation. Some groups are biased. Their opinions are useful for a limited number of users, but not useful for general users and novice users.

Our second approach is the *Diversity Computation*. To estimate the quality or the properties of the groups in social media, the diversity is an important factor. The concept of diversity is actively discussed in social science area to dictate the properties of a group. The current Web is created by a harmonious combination of numerous peoples' activities. Considering this diversity may be a powerful tool for our research to get benefit from the socialized Web.

In this study, we propose three methods for search and analysis of Web contents based on background information. For the social background of links and Web information, we focused on three factors below:

- Why: The reason of links;
  Why did the author of a page create links to a page by editing page?, why did the user of SNS mention a page by reference, or what was the reason to create the link? For instance, they might create links between the current page and the referred page because the referred page was "interesting" or the page was "tear-jerker".

- Who: The property of a person who created the link;
  Individuals and groups have their own properties, which can be estimated by link structure. For instance, the reviewers' credibility must be estimated by using their review history.

- How: Link structure itself;
  Even if the number of links is the same, the properties of a page are sometimes different in terms of the link structure. An obvious example is the diversity; documents linked by similar types of documents and documents linked by diverse documents must have different properties.

First, we focused on the reason *why* the user make a link between SNS sites and the original Web document. We proposed a novel Web search model called "Search-by-Reaction", which allows inputting terms reflecting impression (*e.g.*,"cute" and "tear-jerker") or evaluation terms (*e.g.*, "easy-to-understand" and "useful") as the keyword query. This method enabled the users to search the document such as "the document that everyone evaluate it is cute."

This search model uses the social information, which was not used by existing search models. More concretely, it tries to find and use the reasons why links between Web 1.0 sites and Web 2.0 were established. Web 2.0 sites such as Twitter contain many reactions for the Web 1.0 pages. The reaction is an instance of expression why they read and refer the page. This search model consider the number and the frequency of reactions. We suppose that a page regarded as "interesting" by a lot of people is really an interesting page. However, this supposition brings some problems to our attention. That is, there is no guarantee that a page which gets many reactions containing the term "interesting" is always a truly "interesting" page for everyone. As previously noted, information on the Web 2.0 are made by diverse and many people. To utilize such information as the data source, we must take care of problems caused by social factors.

The credibility of the social media is a big problem to use such resources as a data source to enhance the Web search. The change of the Web usage brings not only richer and more diverse data, but also data that is unreliable. Some of them may be fake, or they may just be low-quality. For instance, if a user who reacts with an "easy-to-understand" to a page about computer is not acquainted enough with computers, his opinion may be wrong. That page may not be actually "easy-to-understand". Moreover, the generality of the opinion is also a big problem. A group whose members have the same interests is a biased group. Their opinion cannot be accepted by common users or novice users. For instance, a page which get reactions like "easy-to-understand" by a group that consists only of computer specialists may be difficult for ordinary people. Since diverse and segmentalized users became able to submit information on the Web, the generality of the information becomes an extremely important factor in our study. Two researches below address the two problems above by using their respective datasets and applications. Both researches use the diversity computation to analyze the credibility and generality of social media data.

The second research topic is solving the problem on credibility. In this research, we proposed a method that focuses on *who* is the evaluator. We estimated the credibility of both groups and individuals by using the diversity-based approach. A dataset of movie reviews was used to analyze the individual reviewers and group of reviewers. This model computed

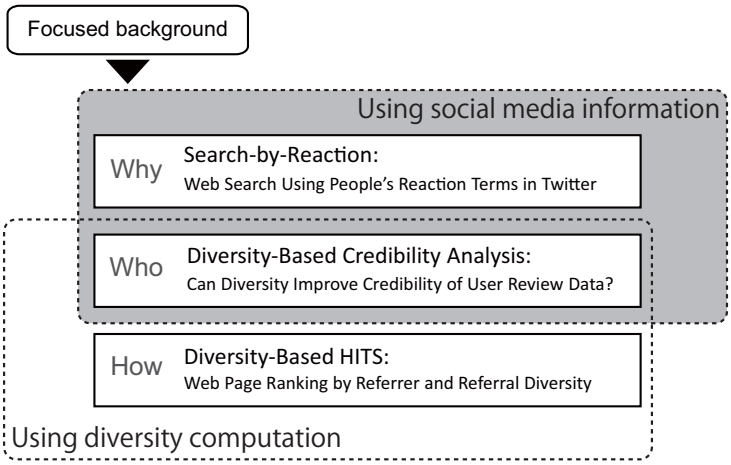

Figure 1.1: Research outline.

the diversity of review histories within groups or individual reviewers, by analyzing links between items and reviewers.

Finally, we focused on the link structure itself: *how* diverse the links are. To solve the problem of the generality, we incorporated the concept of diversity into the mutual recursion of link analysis algorithm. As a dataset, we used a Web graph that contains static Web sites and links between them. Nowadays, it can be said that the Web itself is already social media. We described that to use social data the diversity is necessary above. Even if barriers among social media and the Web have already been eliminated, considering the diversity can help search and analysis of the Web. The current Web has strong aspects of social media. The users who create Web resources are becoming more diverse. Web 2.0 such as CGM, SNS and blogs enable everyone to put the resources on the Web cheaply. From the amateur to the expert, all different kinds of people make Web pages. The web consists of general information and niche information in a harmonious combination. Collaterally, readers or searchers of the Web also become diverse. Finding general information in the Web is important. Our idea to solve this problem is to use the diversity measurements proposed above to enhance recursive link analysis algorithms. Even if the number of links is the same, there result is different if we analyze links between similar documents and links between diverse documents. Our basic idea is that both pages linking diverse pages and pages linked by diverse pages contains general information. We expanded the HITS algorithm [29] with two types of diversity of link structure —the referral diversity and the referrer diversity— that are capable of finding useful documents for novice users.

Figure 1.1 summarizes the aspects and approach that we actually challenged in this study. We managed to use two important techniques.

The first technique is to use social media information to search and analysis. Today's Web is not a set of complete documents, but a mixture of considerable and different information. Thus, it contains not only traditional documents but also logs of real-time communication in the SNSs, product information, user reviews, selling page of e-commerce sites, varied online services like diary, photo storage, and so on. Each service and data source have a different purpose, a different data structure, and a different property. Therefore, the methods we proposed aim to treat these different kind of information in a cross-sectorial manner. The methods refer links between SNS communication and Web documents, links between users and items, and links between all Web sites. Our third research, the diversity-based HITS expands this idea, that is, the Web is already social and we treat Web 1.0 as social media.

The second technique is the diversity computation. As Surowiecki [57] and Page [48] pointed, the diversity is the important aspect to summarize opinions of crowd suitably, or resolve problem by crowds. Nowadays, the Web can reflect social aspects because both its users and its usage is increasing. Diversity is believed to be a helpful tool to handle the data source made from social crowds. Diversity has been actively discussed in the research area of biology and social science. Even in informatics area, diversity has also been discussed, although it was used in a limited sense and a limited aim such as search result diversification. We utilize the diversity more expansively to the information retrieval and information analysis; to utilize diversity to rank Web pages, and to estimate the users' credibility. The measurement of diversity makes it possible to consider the background of links. This idea can be used to solve problems on the generality and on the credibility of our first research.

1.3 Position of My Researches

Figure 1.2 summarizes the position of my researches in the context of the traditional information retrieval area. The most classical paradigm of traditional information retrieval models intended to handle closed documents, such as books, documents, and complete files. Many researchers proposed many methods to retrieve them. These methods rank documents on the basis of the linguistic context. They focus on contents of documents. For instance, the simplest model is based on frequensy of the term appeared in the document, and it was expanded with comparison with another documents like TF-IDF [4, 38], Okapi BM25 [52] and so on. This kind of research is still active topic in recent years, many

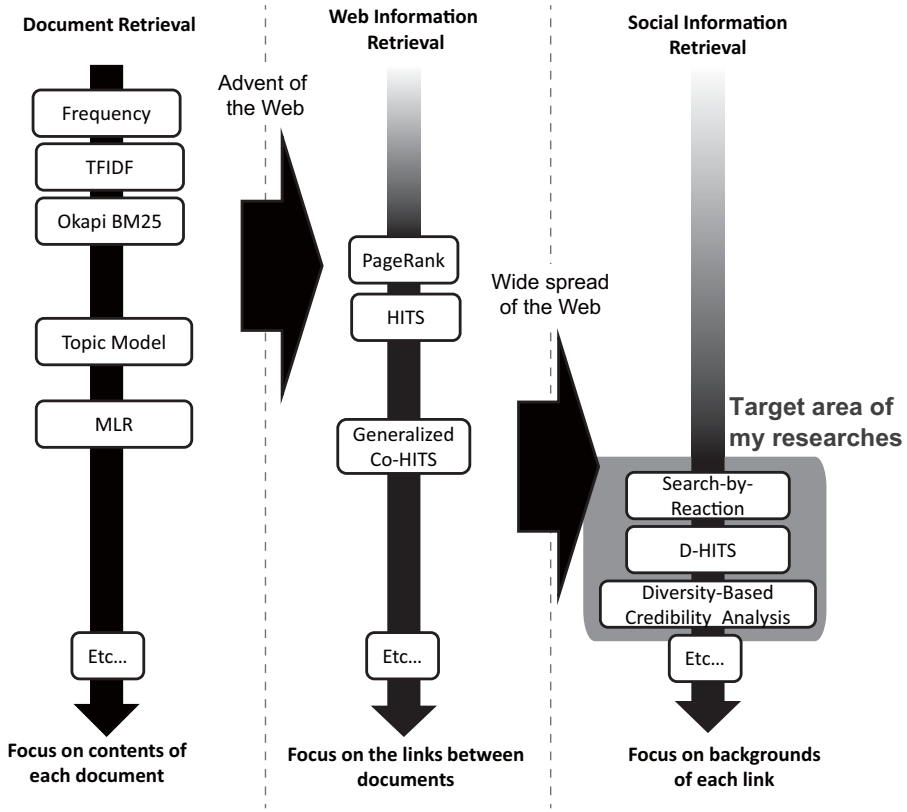

Figure 1.2: Position of my researches.

advanced method were proposed, *e.g.*, topic models based methods by using probabilistic language model [51, 70, 64], machine-learned ranking methods [16, 33].

Meanwhile, the advent of the Web requires the information retrieval methods to embrace new techniques. Documents on the Web are not closed documents but open documents which are linked with each other. They exercise their value by connecting each other closely via the hyperlinks. Therefore, the algorithms based on link analysis became a main stream of Web information retrieval. PageRank [47], the one of the most famous link analysis algorithm, is used in Google to rank Web search result. It calculates the popularity score by the number of links. HITS algorithm [29] is another famous link analysis algorithm. It calculates Hubs and Authorities for each document. Many advanced link analysis methods derive from it, such as Generalized Co-HITS [14], SALSA [32]. These methods focus on the number of links between documents or flow rate of the Web graph to rank Web documents.

Recently, the wide spread of the Web changes the property of Web documents. The number of users, usage and amount of contents on the Web were increasing. People who generate Web documents are becoming diverse, and their aims are diverse as well. Therefore, each document or each link on the Web must have backgrounds: 1) why they create the links, 2) who creates the links, 3) how diverse the link structure is. Two information retrieval paradigms above focus only on the contents of pages and the number of their links. The Web has social aspect today. Thus, it is necessary to consider the social background information of links.

1.4 Thesis Organization

The rest of this thesis is organized as follows:

- Chapter 2
  This chapter briefly reviews the research related to the main methods that are explored in this thesis.

- Chapter 3
  This chapter describes a new concept for improving Web search performance using links between Web 1.0 and Web 2.0 content in a complementary manner. In particular, we describe a social media search, called Search-by-Reaction, which leverages posts on microblogs such as Twitter for the impression- oriented search.

- Chapter 4
  This chapter explains methods to estimate the credibility of reviewers as individuals and as groups. Proposed methods are built on two simple assumptions based on diversity of review history and members of group.

- Chapter 5
  This chapter describes a novel Web ranking method that considers the diversity of linked pages and linking pages. The method uses the resulting diversity scores to expand the basic HITS algorithm to find suitable documents for novice users.

- Chapter 6
  This chapter summarizes this thesis and describe some directions to be explored in the future.

CHAPTER 2

Related Work

This study tackles with the problem of link analysis by considering background information. The objective of our method is to find Web documents or users by novel information needs, such as suitable documents for impressions, reliable users or reliable groups and suitable documents for novice users. This chapter introduces past studies on link analysis algorithms and the diversity, as well as the research on information needs related to our work.

2.1 Link Analysis Algorithms

In this thesis, we proposed a novel link-based algorithm. Link-based approaches have already been explored widely in the past. There are many research work that extend the PageRank [47] and HITS [29] from several aspects.

The topic sensitive PageRank [19] deals with the topic of the query and documents. The TrustRank [18] algorithm uses PageRank for spam filtering on the basis of the simple theory that spam pages link to both good and spam pages but good pages link only to good pages. Another idea is to involve the concept of time and space into PageRank to measure the impact of historical events and entities [58].

The HITS algorithm [29] is an alternative approaches featuring Web graph analysis uses the link structure on the Web to locate communities in the Web graph. It models the Web graph as a bipartite graph and calculates the importance of Web pages by convergence calculation. It uses two types of scores: a Hub score and an Authority score. Authorities are pages that show valiant information. A page obtains a higher Authority score if it is linked by more pages obtaining high Hub score. Hubs are pages that link to good Authority sites. A typical example of a good Hub page is a linking site or a good search

result page. The page gets higher Hub score if it is linking more pages obtaining high Authority score. The Generalized Co-HITS algorithm [14] is a HITS-based algorithm that extends the conventional HITS algorithm from Web links to general bipartite graphs such as a paper-and-author pair. SALSA [32] is an similar algorithm, which makes a bipartite graphs based on Hubs and Authorities and calculates scores for nodes on the graph by applying random walk. Fast random walk with restart [59] is similar, too, and is used to compute the similarity between nodes.

Our algorithm is a weighted HITS-based algorithm. The unique part is that it defines the weight of the edges depending on the diversity.

In this study, the proposed methods not only utilize links between Web pages, but also leverage cross domain links between different kinds of sources, such as social networking services, movie review services. They use the links between tweets and Web pages, the links between users and items. Many studies on information retrieval have focused on using external information to complement the main content of the document. There are a lot of resources that can be used for Web searches, including Wikipedia [22], community QA sites [62], blogs, and social annotations.

A typical example of social annotations is social bookmarking [20]. In such services, users can annotate Web pages with tags and share with others. Users can search for documents as if they are classified by their taxonomy. This situation is called folksonomy. Some studies have approached in that way to enhance Web searches [21, 23, 63, 9, 6]. Social bookmarking services are emerging Web services that help users to share, classify, and discover interesting resources. Yanbe *et al.* [66, 67] explored the concept of an enhanced search in which data from a social bookmarking system is used on the Web. They proposed to combine the widely used link-based ranking metric with the one derived using social bookmarking data. For instance, it adds "freshness" and "user-interestingness" as ranking metric. Additionally, tags in Consumer Generated Media are useful resources to enhance the Web document retrieval. Kato *et al.* [25] developed a method that replaces an abstract query term given by a user with a set of concrete terms and that uses these concrete terms in queries as inputs to search in conventional image search engines. The method proposed in this thesis uses reactions in Twitter as if they were tags. However, while tags consist of a few terms, reactions are written in natural language. Another difference is that tags usually express explicit impressions. However reactions are posted usually after reading some impressive document as a by-product of communications. Reactions are implicit impressions. In addition, tags have limited diverseness and dimensions. Therefore, reactions show more complexity than the tags. It may have a potential to accommodate

specific information needs.

Using Web communication data as knowledge base is an active research topic, which can be applied for improving the Web search. In particular, a major discovery is that one can easily find trends or follow events from Twitter feeds [13, 53]. Dong *et al.* [15] measured the freshness of Web sites in terms of the frequency of URI citations in Twitter. Freshness scores were used to improve Web search results. The Web page in touch with a "hot" topic is higher ranked in the search results. Social factors in Web communication can be used for enhancing Web searches [55]. Smyth *et al.* [41] have made available a social search system named *HeyStack*[*], which blends Web search engines and Web communications. Lu *et al.* [36] use SNS data to estimate the adequacy of reviews and remove untrustworthy reviews. Relationships among the users on Web communication sites are useful for personalizing search results [12]. Google started Google Social Search[†] in 2010 as personalized Web search engine which uses friendship relation information. It ranks pages juried more positive by more friends of the searcher at higher positions.

Other forms of Web 2.0 contents are also being used to enhance Web searches [62]. Wikipedia is a good external resource to enhance Web search results [22]. In particular, Hu *et al.* [22] used Wikipedia category information to categorize user queries.

2.2 Measuring Diversity

Our proposed method incorporates a diversity-based measure to find experts and evaluate the credibility of a group of reviewers. There have been various studies on diversity specialized for different problems.

Collective intelligence has been actively discussed, as the collaboration on Web sites became a popular activity. Surowiecki [57] presented in his book some conditions of data under which the wisdom of crowds works correctly: *diversity of opinion*, *independence*, and *decentralization*. Once the three requirements are satisfied, useful knowledge can be built from the data by means of *aggregation*.

Diversity has been extensively used in the field of information retrieval. One of the most active research topics is diversification of Web search results [1, 61, 10]. For example, maximal marginal relevance [11] was used to diversify search results by decreasing the score of the pages similar to ones ranked in higher positions. The aim of our research is not to diversify search results, but to use diversity as the feature of ranking.

The research areas that focus on diversity are not limited to computer sciences, but

[*]http://www.heystaks.com/

[†]http://googleblog.blogspot.com/2010/01/search-is-getting-more-social.html

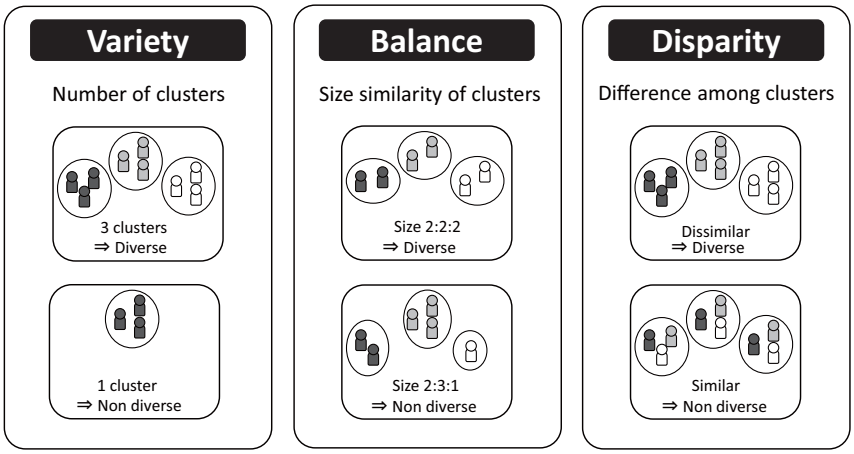

Figure 2.1: Three factors of the diversity.

include sociology, ecology, life science, economics, *etc.* Stirling [56] summarized three key factors regarding categorical diversity: *variety*, *balance*, and *disparity*. Figure 2.1 shows the basic concept of these three factors. Many diversity measures have been proposed especially in the biology area. Biodiversity has recently received attention, and is measured by Shannon-Wiener index [46], which was developed based on Shannon entropy. The index highly correlates to the number of breeds and the balance across different breeds. Another diversity index measure called Simpson's diversity index [37], is defined as the probability of breed coincidence of two randomly-selected individuals.

Since the diversity is a multi-faceted concept as can be seen in the earlier discussion, the optimal design of a diversity measure highly depends on its application domain. In this thesis, we use two different kinds of diversity measures for reviewer groups, namely, entropy-based and variance-based diversity measures. The former measures the variety and balance, while the latter measures the disparity of reviewers. These two measures were compared in our experiments.

2.3 Related Information Needs

2.3.1 Impression-Based Retrieval

Opinion mining and sentiment analysis related to impression-oriented information needs that the Search-by-Reaction aims to accommodate. There are many research that mine reputations or opinions from Web content, especially Web 2.0 contents [30, 28]. They search

for social reputations from BBSs, blogs, SNSs, and so on. The reputations of products (*i.e.*, what buyers feel about them) are retrieved. Opinions from their about products are similar to impressions about Web sites. We have a different goal from those research. What this thesis proposes is a Web search model, which takes an impression as its input and search result ranking which consists of Web sites as its output. In contrast, most opinion mining systems aim to find reputations by using the name of a product as a query. Products have standpoints and attributes determined by a class of products. An opinion is calculated as a set of attributes and sentiments [24]. For example, each TV has attributes like "size", "picture quality", "price", and so on. Sentiments, which are either positive or negative values, are calculated for each attribute. The sentence "It is inexpensive" has a positive sentiment, and its attribute is "price". However, Web pages have unnumbered attributes. Systems cannot determine attributes preliminarily. Thus, it is hard to guess what the user input as a query. These methods, which use fixed attributes are not suitable for Web search. Our method does not need a stationary dictionary of attributes.

Some research focuses on readers' sentiments about documents [50, 26, 71]. Sentiment analysis estimates how a reader feels after reading a document by using terms in the main contents of the document. It calculates a sentiment for every term in the document. Emotionally-charged terms make readers emotional. For example, the term "rain" would have a negative sentiment. These methods guess the readers' emotions by using only terms included in the document. However, readers do not always feel and express the same as writers do. Therefore, there always exists an expression or terminology gap between readers and writers, which results in the difficulty of search.

It has become popular to mine sentiments from Twitter [35, 49, 8, 7, 17]. A sentiment is calculated for each tweet. Mined sentiments are used as reputations of the target as far as Twitter users are concerned. Sometimes, sentiments are used to enhance Web searches [27, 43]. In particular, it can be used to search for non-text content such as movies [44, 72], music [65], and images [31]. Such multimedia contents cannot be obtained by using a query term by means of text matching. These systems connect the reader's sentiment with features of content. Sentiment is determined by score propagation. The seed of propagation is a few pieces of content that the user annotates as the query.

2.3.2 Expert Finding

The research problem on finding experts has a long history [68, 40, 5]. Finding experts from consumer generated media (CGM) sites has been an active research area in recent years. One of the representative examples is expert finding in community-based question

and answering (cQA) sites. Liu and Koll [34] proposed a method to find experts from cQA sites by focusing on the past answers given by users. In this work, experts are defined as users who can answer a certain kind of questions. The basic assumption used in their method is that users are able to answer a question if they have answered similar questions in the past.

There is some previous work on discovering experts to improve the accuracy of recommendations. One of the assumptions in this line of work is that an item evaluated as high-quality by experts is likely to be high-quality for many other users. Amatriain *et al.* [3] proposed a recommendation method that utilizes only the *nearest experts*, which are defined as users who posted a sufficient number of reviews, and are the most similar to a user who receives a recommendation. The performance of the proposed method was comparable to traditional collaborative filtering algorithms, even when a small expert set was used. Their expert detection method was based solely on the number of reviews, and the method did not take into account reviewed items. In our work, we utilize the diversity of reviewed items to find experts, and propose a method to aggregate reviews to precisely estimate the quality of items.

Sha *et al.* [54] proposed a method of seeking two different kinds of experts from an online photo sharing community: *trend makers* and *trend spotters*, and recommending trends in the community estimated by these experts.

McAuley and Leskovec [39] proposed a method to find domain experts by using their review experience. Users are expected to become more professional in a domain if they work on the domain for a longer time. This work pointed out two important perspectives of expertise: 1) a user becomes an expert if s/he has been engaged in a domain for a long time, and 2) the evaluations done by novices tends to be diverse, while those by experts tends to be focused.

2.3.3 Finding Information for Novice Users

There also exists a lot of research work exploring to estimate the specialty of a document by an approach of content analysis. In particular, estimating the specialty of terms contained in documents is a hot topic.

In the field of natural language processing, several methods that extract the special terms (*e.g.*, technical terminology and jargon) from documents have been proposed. Nakatani *et al.* [45] proposed a link analytic method using the Wikipedia category structure to extract special terms. These studies used big corpora or document sets of limited specialized domains and structural information to measure the specialty of a term. Our method

does not aim to find special terms but rather special Web pages without using particular datasets. The existing methods can be used to increase the accuracy of our method in a complementary style. There is a previous similar study that aims to find comprehensible Web pages by the link analytic approach. Akamatsu *et al.* [2] proposed a TrustRank-based method built on one simple rule: comprehensible pages are more likely to link comprehensible pages. General pages that we want to find are similar to comprehensible pages, but this method is based on TrustRank, which focuses only on the number of links and not on how they link.

CHAPTER 3

Search-by-Reaction

3.1 Introduction

Conventional Web search engines always suffer from low precision and recall, especially when the contain terms related to users' impressions and evaluations. We propose an approach to enhance the performance of the conventional Web search engines by allowing users to directly search with impression and evaluation terms usually posted on micro-blogs. Web 2.0 consumer generated media (CGM), such as social bookmarks, micro-blogs, blogs, community QA (cQA) sites, and Wikipedia, prove to be valuable and important collections of human knowledge created by users in a collaborative manner.

For example, Wikipedia is a user-generated encyclopedia of *collective knowledge* on the Web. Blogs and micro-blogs (Twitter* is a typical example) contain a large volume of user-centered information on their experiences and viewpoints. Furthermore, cQA sites onthe Web provide a large amount of collective knowledge consisting of questions and answers generated by users. Indeed, cQA sites contain a lot of direct requests for information and solutions to the problems.

However, queries that effectively work for the conventional Web search engines are confined to these containing terms that appear in the contents of desired documents since these engines search for documents by exactly matching the keywords in queries. This kind of mechanism highly impacts on the performance of the search engine, resulting in low precision and recall especially when users formulate queries with their own impressions and evaluations regarding topic, which are unlikely to appear in target documents. For

*https://twitter.com/

example, conventional search engines do not work well for queries containing subjective evaluation terms such as "good" or "useful".One typical type of useful Web contents (*e.g.*, news articles, online dictionary pages, recipe and so on) contain very few of such subjective evaluation terms. On the other hand, many spam pages or commercial sites are likely to contain many evaluation and impression terms to publicize their products. All the reasons mentioned above make the search engines difficult to return users documents related to the queries with impression or evaluation terms. As is known, however, Web 2.0 communication data contain valuable information reflecting users' viewpoints. Such Web 2.0 data may help to fill this gap by connecting users' impression and evaluation terms to users' reactions on Web 2.0 communication sites, so as to support a wider class of query terms and further increase the precision and recall of conventional Web search engines. In this chapter, we propose a method to enhance Web search engines by enabling the search engine to accept impression (*e.g.*, "cute") and/or evaluation (*e.g.*, "useful") terms as query keywords by leveraging people's reactions (*e.g.*, phrases such as "it's convincing") posted on micro-blogs. The novel features of our system are summarized as follows.

- **Reader-based information retrieval model:**
  Conventional Web search engines are based on the *writer-based* information retrieval model. That is, search users are forced to formulate search queries by guessing terms that will frequently appear in their desired documents (retrieved Web pages). Indeed, search users cannot get high-precision search results for query terms that represent how people "feel" about the documents. Our system allows users to input readers' impressions or evaluations as query keywords. In this sense, our system supports a *reader-based* information retrieval model.

- **Association between reaction terms and input impression or evaluation terms:**
  Our system discovers and uses people's *reaction terms* in Twitter to find relevant Web pages associated with user impression and evaluation terms in queries. Indeed, Web communication data, such as tweets in Twitter, often contains references to Web pages and users' reactions to the referred Web pages.

- **Improving Web 1.0 search performance by knowledge extracted from Web 2.0:**
  Our idea of improving Web 1.0 search performance by using external knowledge extracted from Web 2.0 is illustrated in Figure 3.1. The system accepts queries that contain non-conventional intents, *e.g.*, the impressions or evaluations. The system

uses social information in Web 2.0 to accommodate these queries. In this research, we focus on the opinion of users in Twitter.

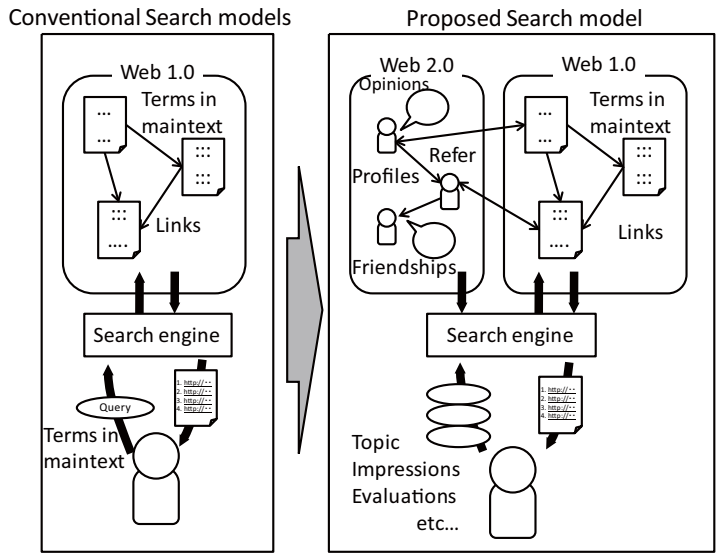

Figure 3.1: Enhancing Web 1.0 search by using Web 2.0.

3.2 Methods

The Search-by-Reaction is a search model that accepts an impression term or an evaluation term as a part of query. It investigates users' reactions posted on Web 2.0 contents as annotations.

Conventional Web search engines cannot find "easy-to-understand documents about C-language" by a given query "easy-to-understand C-language" because the search engine only knows the terms appearing on documents. Impression terms or evaluation terms do not always appear in the documents. In addition, even if the documents contain the term "easy-to-understand", it does not mean that they are actually easy to understand.

Our system focuses on reactions in Twitter, one of the most typical Web 2.0 communication site. In Twitter, it is common that someone recommends a Web site to his/her friend, and in return, the friend posts a tweet containing a reaction like "It's easy to understand!!." Our system uses such reactions as annotations of readers' impressions for recommended Web pages.

Figure 3.2 shows the outline of our system. The input is a pair of keyword queries:

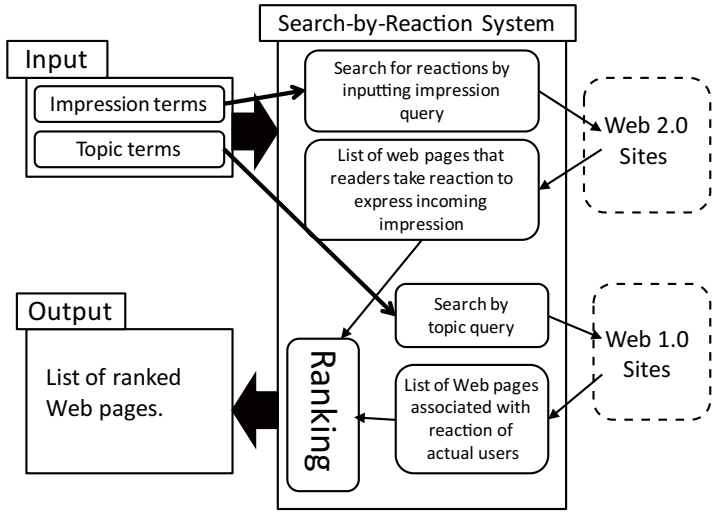

Figure 3.2: Outline of the Search-by-Reaction system.

an *impression query* and a *topic query*. An impression query is a term that expresses the impression that a user expects to feel after reading the documents. A topic query is a term that expresses what the retrieved documents are about (*e.g.*, a traditional Web search engine query). The output is a ranked list of Web pages. The system uses Web communication data as well as the contents of the pages to rank the result for given impression queries. The relevance between Web pages and impression queries is calculated by using users' reactions to the pages posted on the Web communication sites. The relevance between the Web page and the topic query is calculated from the content of the retrieved Web page. For example, when a given query is ("cute", "cat"), the system searches for Web pages that contain "cat" in their text and a reaction for these pages such as "I want to hug it!!" which has similar meaning to "cute" posted on Twitter.

The impression query is the term indicating readers' feeling after reading a document. When a user comments on a document that it is "cute", it usually does not literally mean that "This page is cute." However, the user may desire to convey the idea of "Wow, I want to hug this" in reference to the some specific contents of the document thought to be cute. It is common that the actual reaction does not include a corresponding impression term. The system has to consider how reactions are semantically similar with the given impression query. There are three types of reactions that express the same impressions. We explain each of them by taking an example in which a user is looking for an "easy-to-understand" Web page.

- **Similar meaning terms:**
  A very simple paraphrase or synonym: *e.g.*, "It's not so difficult," "Comprehensible!."

- **Paraphrasing based on cause and effect:**
  An outcoming of easy-to-understand documents or why these documents are easy-to-understand. *e.g.*, "It's written in detail how I should do!" or "Thank you! That solved my problem!."

- **Exclamatory expression:**
  A simple expression or phrase about how the reader feels: *e.g.*, "Aha!" or "I get it!."

The system uses these different types of reactions to generate the ranking that contain pages which are taken reactions semantically similar to given impression queries. To this end, we propose a measure that calculates semantic similarity between given impression terms in queries and terms used in reactions.

3.2.1 IR Model

Each query $q = (q_{\mathrm{I}}, q_{\mathrm{T}})$ consists of an impression query $q_{\mathrm{I}} \subset V$ and a topic query $q_{\mathrm{T}} \subset V$, where V is a set of all terms. A ranking function $rank(p, q_{\mathrm{I}}, q_{\mathrm{T}})$ is defined as follows:

$$rank(p, q_{\mathrm{I}}, q_{\mathrm{T}}) = score_{\mathrm{T}}(p, q_{\mathrm{T}})^{\alpha} \cdot score_{\mathrm{I}}(p, q_{\mathrm{I}})^{\beta}, \tag{3.1}$$

where $p \in P$ is the Web page and the P is the set of all Web pages. The topic score $score_{\mathrm{T}}(p, q_{\mathrm{T}})$ is the relevance of the Web page to the topic query q_{T}. The impression score $score_{\mathrm{I}}(p, q_{\mathrm{I}})$ is the relevance of the Web page to the impression query q_{I}. α and β work as weight for each score.

3.2.2 Topic-Score

The topic score $score_{\mathrm{T}}(p, q_{\mathrm{T}})$ means the relevance between a topic query q_{T} and a Web page p. It is the same as the relevance score of conventional Web search engines. The topic score $score_{\mathrm{T}}(p, q_{\mathrm{T}})$ is estimated by using the terms occurring in the main text of p. This model uses a simple query likelihood [38] without smoothing:

$$score_{\mathrm{T}}(p, q_{\mathrm{T}}) = \Pr(p|q_{\mathrm{T}}) \propto \Pr(q_{\mathrm{T}}|M_p), \tag{3.2}$$

where M_p is the language model that generates p. We build this model on the basis of the uni-gram fo words appearing in p. The function $\Pr(q_{\mathrm{T}}|M_p)$ is calculated by as:

$$\Pr(q_{\mathrm{T}}|M_p) = \prod_{t \in q_{\mathrm{T}}} \Pr_{\mathrm{mle}}(t|M_p) = \prod_{t \in q_{\mathrm{T}}} \frac{f_{t,p}}{L_p}, \tag{3.3}$$

where $f_{t,p}$ is the term frequency at which a term t appears in q_{T}, and L_p is the number of terms appearing in p.

3.2.3 Impression-Score

The impression-score $score_{\mathrm{I}}(p, q_{\mathrm{I}})$ is a relevance between a topic query q_{I} and a Web page p.

3.2.3.1 The Simplest Reaction Search

In the simplest case, we can calculate the $score_{\mathrm{I}}(p, q_{\mathrm{I}})$ as how many times q_{I} appeared in the reactions of p as bellow:

$$score_{\mathrm{I}}(p, q_{\mathrm{I}}) = |\{r | r \in R(p) \wedge q_{\mathrm{I}} \subset r\}|, \tag{3.4}$$

where $q_{\mathrm{I}} \subset r$ means reactions which contain q_{I}, and $R(p)$ means the reactions of p.

One of the major problem of this straightforward method is that when users post their reactions to the recommended Web pages on Web communication sites, such reactions are usually written in natural language. Sometimes, the reactions may contain grammatically ambiguous sentences or oral expression, and facial expression marks (emoticons). Moreover, readers often express their impression in language that avoids any specific impression. For example, when someone reads a "cute" Web page about a cat, their reaction might be "I wanna pet it!!" or "fluffy :-D" rather than "It's a cute cat." These reactions express the same impression, "cute", but with different denotative meanings. To measure relevance between impression queries and reactions more accurately, we need to calculate the semantic similarity between them.

3.2.3.2 Finding Another Expression of Impression Query

We present a simple yet reasonable algorithm to find reaction term (*e.g.*, "hug", "fluffy") rlated to the input query q_{I} (*e.g.*, "cute"). The basic idea behind this algorithm is illustrated in Figure 3.3. In this example, the system searches for reaction terms related to the impression "cute". Each of p_1, p_2, p_3 is a Web page, and $r_1, r_2, \cdots, r_7$ are the reactions for these pages. Each reaction consists of several reaction terms. The basic rule is very simple; it is similar to TF-IDF. The system searches for terms included in only the reactions for Web pages evaluated by someone as q_{I}, *i.e.*, p_1 and p_2, which the users take reactions including "cute". In this way, the term "pretty" regarded as being another expression of "cute" because it appeared twice in reactions of "cute" pages, and never appeared in these without "cute". The procedure is described in detail as below.

Web pages sometimes recieve multiple reactions from several people, not from a single

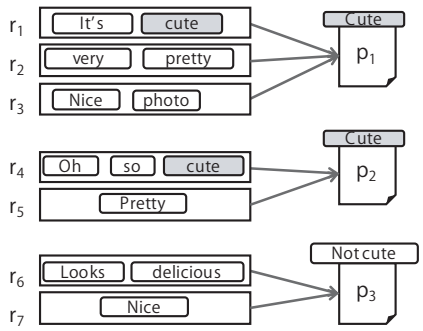

Figure 3.3: Simple example of calculating the impression score.

person. To calculate the relevance $score_{\mathrm{I}}(p, q_{\mathrm{I}})$ between p and q_{I}, the system need to calculate scores of these reactions. We achieve this by calculating the arithmetic mean of the relevance score $score_{\mathrm{I}}(p, q_{\mathrm{I}})$ for each reaction to q_{I} as follows:

$$score_{\mathrm{I}}(p, q_{\mathrm{I}}) = \frac{\sum_{r \in R(p)} s_r(r, q_{\mathrm{I}})}{|R(p)|}. \tag{3.5}$$

Therefore, the reaction r is a natural language sentence, and it consists of several terms. The relevance score for a reaction r is calculated by comparing each term in r to q_{I}. For example, when the reaction r is "Wow! So pretty :-)," the system calculates the relevance to q_{I} for each "Wow!", "So", "pretty" and ":-)". In this way, the relavance score $s_r(r, q_{\mathrm{I}})$ of an individual reaction is calculated by

$$s_r(r, q_{\mathrm{I}}) = \frac{\sum_{w_j \in W(r)} s_w(w_j, q_{\mathrm{I}})}{|W(r)|}, \tag{3.6}$$

where $s_w(w_j, q_{\mathrm{I}})$ is the relevance score of each term w_j appearing in $W(r)$, which donates all terms included in r.

We calculate the score $s_w(w_j, q_{\mathrm{I}})$ is as how the term w_j expresses a nearly identical meaning impression from q_{I}. Specifically, $s_w(w_j, q_{\mathrm{I}})$ is defined by how characteristically w_j co-occur with q_{I}:

$$s_w(w_j, q_{\mathrm{I}}) = |P_{\mathrm{reaction}}(q_{\mathrm{I}}) \cap P_{\mathrm{reaction}}(w_j)| \cdot \frac{1}{|\{r | w_j \in r\}|}, \tag{3.7}$$

where $P_{\mathrm{reaction}}(w_j)$ is the set of Web pages that have reactions including the term w_j. The first term of $s_w(w_j, q_{\mathrm{I}})$ increases when w_j were included in reactions to the most pages that also have reactions of q_{I}.

Say, terms focusing on a similar aspect with "cute", such as "hug", would appear in reactions to a page in which it might contain a cute furniture, cute soft toy, cute animal photo page, and so forth, whereas the term "useful" would be included only in reactions to the page with cute furniture but not with other cute pages. The second term of $s_w(w_j, q_\mathrm{I})$ decreases if w_j has a wide range of meanings or be a common term. The value of $s_w(w_j, q_\mathrm{I})$ drops from zero to one since all Web pages in our dataset have one or more reaction.

3.3 Experiments

We built a Web application embodying the above method and recruited people to make a answer set by judging every result given by the system. We conducted two experiments. Section 3.3.3 reports the experiment for the algorithm fo finding another expression terms, which is proposed in Section 3.2.3. Section 3.3.4 reports the experiment for the ranking algorithm proposed in Section 3.2.

3.3.1 System Implementation

Figure 3.4 shows a screenshot of the system. This system interface contains two text boxes that receive input queries; the left one accepts an *impression query*, and the right one accepts a *topic query*. Output is a ranked Web page list resolved in accordance with the ranking function explained in Section 3.2. Each Web page is shown with reactions and a short snippet (Figure 3.5).

Reactions to the Web pages are taken from Twitter. We focused on reprinting (called *retweets*) in Twitter. Users often retweet someone else's tweet while adding their own comment. These comments are then used as reactions. Our system regards the reprinters' comments as reactions and the URIs contained in retweets as the target pages for these reactions. Figure 3.6 shows an example of a reaction. In this example, "Looks so delicious! I wanna go!" is the reaction, whereas "http://r.tabelog.com/kyoto/A2603/A260302/26006618/" is the URI of the Web page to which the reaction is annotated.

The system flow of the Search-by-Reaction search engine is as follows:

1. Collect reactions and URIs from Twitter. The data set is stored in database.

2. A user inputs a impression query and a topic query to the text boxes.

3. Determine the list of word scores $s_w(w_j, q_\mathrm{I})$ by using the scheme explained in Section 3.2.3.

Figure 3.4: Screenshot of the system.

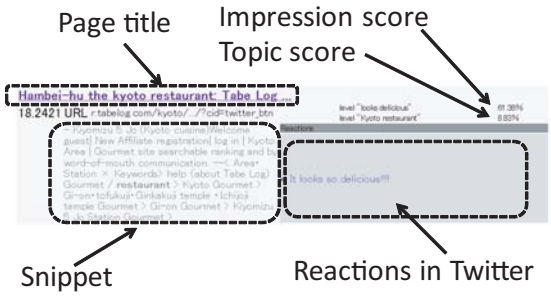

Figure 3.5: Sample results page.

Figure 3.6: Example of reactions.

4. Search for Web pages with reactions to the topic query from DB.

5. Calculate the impression scores by using score of each term calculated in step 3

6. Calculate the topic scores by main text for each page.

7. Make an integrated ranking using the impression scores and the topic scores.

3.3.2 Test Set

The system used Twitter streaming API[†]. Reactions were collected from November 2010 to December 2010. The number of tweets crawled amounted to 30,134,604. All of them were written in Japanese, and URIs are included. We found 2,370,222 tweets that were in the prescribed format. The system used these tweets as reactions for Web pages. There were 1,286,752 non-duplicate URIs after decoding. These pages are regarded as documents to search by the system.

3.3.3 Same Meaning Reaction Evaluation

Experiments

The method of finding another expressions of a given impression query is similar to that of query expansion. We compared the proposed method with the existing query expansion method such as pseudo relevance feedback (PRF) [69]. Two participants evaluated the top 30 terms generated by both the proposed method and PRF for ten queries.

The PRF method ranks Web pages according to Equation 3.8 and regard the top K results as relevant pages. Terms in reactions to these pages are used as expanded queries.

$$score_{\mathrm{PRF}}(p_k) = \left| R_{\mathrm{include}}(q_{\mathrm{I}}) \cap R(p) \right|, \tag{3.8}$$

[†]http://stream.twitter.com/1/statuses/

Proposed	PRF
poignant	poignant
factory	good
lacrimal	not
feel-something	cried
can-cry	tear
goose-bumpy	T (part of emoticon)
cry	great
ululant	excellent
again	but
cry	want-to

Table 3.1: Expanded terms for "poignant".

where $R_{\text{include}}(q_{\text{I}})$ is a set of reactions which contains q_{I}. Score of each expanded query is defined how many times the term w_j appeared (see Equation 3.9).

$$s_{\text{PRF}w}(w_j) = \sum_{k=1}^{K} |R(p_k) \cap R_{\text{include}}(w_j)|. \tag{3.9}$$

We set $K = 10$. To evaluate the effectiveness of proposed methods, we used the precision based metrics such as P@10 and P@30 (precision at 10 and 30), and Mean Average Precision (MAP) [38]. Table 3.1 shows an examples of expanded terms suggested by the proposed method and PRF method for the term "poignant". Two participants evaluated each expanded terms.

Results

Table 3.2 shows the overview of the result. Since the proposed method achieved higher scores on MAP, the top 30 ranking of proposed method performed better than the ranking of PRF through all the tested queries. The proposed method performed the baseline at P@10 and P@30 for almost tested queries. As exceptional examples, baseline method won when the queries were q_{I}="interesting" and q_{I}="excellent". When q_{I}="interesting", the proposed method returned terms such as "challenge", "idea", "view-point". These terms co-occurred with "interesting" in many cases, but did not represents same impression with the query. The terms, "laugh", "ha-ha", "lol" and so on, are evaluated by participants to have same impression with the term "interesting", and these terms are used very frequently in Twitter. When the query was "excellent", both methods were unable to get suitable

	Proposed		PRF	
q_I	P@10	P@30	P@10	P@30
aha	0.60	0.33	0.10	0.03
assent	0.20	0.10	0.00	0.00
easy	0.20	0.07	0.00	0.00
tear-jerker	0.80	0.67	0.60	0.41
interesting	0.50	0.50	0.60	0.31
poignant	0.70	0.47	0.30	0.31
feel-fear	1.00	0.60	0.10	0.03
surprising	0.10	0.03	0.00	0.00
excellent	0.10	0.07	0.30	0.10
awful	0.60	0.30	0.20	0.17
MAP	0.73		0.31	

Table 3.2: Score of each term.

terms, since the term "excellent" has rather broad meanings.

3.3.4 Ranking Evaluation

Experiments

Our system was compared with two baseline search engines.

- **Baseline 1**: AND search
  This search engine is the same as conventional Web search engines. It accepts an impression query and topic query, but both of these queries are treated in a same way. It scores Web pages by checking if the query terms appear in the main text of a page, so that pages which include both the impression query and topic query were retrieved. The score of the retrieved document in the baseline system is explained in Section 3.2.2.

- **Baseline 2**: Simple Reaction Search
  The other baseline system is the simple reaction search described in Section 3.2.3.1. It uses readers' reactions to find relevant pages but it does not count reactions written by another terms from q_I.

We prepared 28 pairs of topic and impression queries (see Table 3.3), and three systems ranked documents by using these queries. The top 20 results were evaluated on the ba-

sis of the relevance by eight examinees. The pages were shuffled for fair comparison, and examinees scored each pages by two relevance measures: impression relevance and topic relevance. The impression relevance score was labeled in five-point scales. The topic relevance score is one or zero. The maximum number of search results generated by proposed system was limited in 5,000 in the experiment. Thus, we ignored pages ranked lower than 5,000 when calculating impression scores. The number of search results of baseline 1 was limited in 3,000. To evaluate the effectiveness of the impression-based search, we set the lower weight for the topic score $\alpha = 0.3$ and the higher weight for the impression score $\beta = 1.0$.

P@k shows how many relevant pages are ranked in the top k results. Column # of pages in Table 3.2 denotes as the number of the pages retrieved by each system. To consider the reason why the page was evaluated as irrelevant, two relevance score, *i.e.*, topic relevance and to impression relevance were treated separately. Topic relevance of each page become one if a majority of examinees evaluate it as one. We regard pages as relevant in terms of impression if and only if the average of the impression relevance from examinees is higher than three. Therefore, MAP were computed after topically irrelevant pages were removed.

Result

Table 3.3 shows the results for the each metric. The column # of pages including q_{I} shows the number of pages retrieved by the proposed method and that include q_{I} in their main text. The column # of new relevant pages denotes the number of pages retrieved by the proposed method that did not contain q_{I} in main text, that is, conventional search engines could not find them. Our system outperformed the baseline systems not only on MAP, but also on the average of number of finding pages at P@k over all the queries (see Table 3.4). The average of P@20 of our system was significantly higher than that of baseline systems by student's t-test. The p-value between the proposed method and the baseline is 0.03, and one between proposed method and the baseline 2 is 0.00.

3.4 Discussion

Larger number of relevant pages were retrieved than two baselines. Our system found a lot of pages that are not found by the traditional Web search method since those pages do not contain impression terms in their content. However, even if the document does not contain impression terms in its content, the proposed system can finally find the document. If the document contains only topic terms, the relevance between the document and impression

		# of pages			# of pages	# of new	P@20		
q_{I}	q_{T}	baseline1	baseline2	proposed	including q_{I}	relevant pages	baseline1	baseline2	proposed
aha	cold mechanism	52	6	215	3	3	0.00	0.10	0.15
assent	joblessness matter	93	9	457	3	5	0.10	0.05	0.25
aha	word-origin	20	6	190	4	12	0.15	0.25	0.70
easy-to-understand	cloud-computing	331	13	*4,031	1	5	0.10	0.05	0.25
easy	diet	*3,000	28	3,847	13	2	0.10	0.00	0.10
informative	How-to-study	53	6	762	1	7	0.35	0.10	0.40
tear-jerker	date diary	54	6	1,237	3	0	0.00	0.00	0.00
interesting	gate diary	346	13	1,231	7	0	0.00	0.00	0.00
impressive	dog memory	437	10	1,238	5	1	0.20	0.00	0.25
tear-jerker	exam diary	45	6	1,434	3	0	0.05	0.10	0.10
feel-fear	clinic episode	455	35	1,827	10	0	0.10	0.00	0.00
gammy	junior-high memory	240	9	713	7	0	0.15	0.00	0.00
cute	dog picture	*3,000	290	*4,001	11	7	0.20	0.35	0.35
feel-aggravated	news arrest	219	6	*3,699	2	2	0.05	0.05	0.10
surprising	show-business-news	964	14	*3,506	4	5	0.30	0.05	0.25
excellent	NewYear's-card	745	135	*4,006	6	4	0.20	0.40	0.20
surprising	news football	747	18	*4,063	0	3	0.00	0.10	0.15
want-to-try	recipe tomato	47	7	670	2	0	0.00	0.00	0.05
want-to-drink	cocktail	7	8	1,228	1	10	0.15	0.20	0.55
interesting	mystery novel	124	8	321	8	5	0.20	0.05	0.30
looks-delicious	noodle Kyoto	19	5	1,085	0	4	0.00	0.05	0.20
want-to-go	restaurant Kyoto	51	19	1,839	12	10	0.15	0.20	0.50
long-forgotten	game RPG	357	37	*3,904	5	1	0.15	0.20	0.15
interesting	action-movie	58	5	187	3	5	0.30	0.00	0.25
want-to-get	novelty	641	85	2,505	4	2	0.15	0.10	0.15
cute	cat photo	*3,000	526	*4,038	8	12	0.95	0.60	0.60
bad-looking	cat photo	14	1	226	0	0	0.10	0.00	0.00
awful	dog pet photo	223	15	1,666	6	1	0.10	0.00	0.10

Table 3.3: Queries and results.

Metric	baseline1	baseline2	proposed
MAP	0.57	0.56	0.60
P@20	0.15	0.11	0.22
P@10	0.18	0.16	0.24
P@5	0.26	0.25	0.29
p-value(P@20)	0.03	0.00	–

Table 3.4: Results through all queries.

query q_I is calculated by using its reactions, and it increased the number of the retrieved document. Although the number of retrieved pages increased, the precision was improved far from degraded. There are many pages that include impression terms but they do not leave the impression for readers. However, there existed some pages which contained impression terms, but not the desired documents the users expected. In this case, our system correctly ranked these pages in a lower rank. The baseline 2 seems to work well on the precision-based metrics, however this method can only access very few related pages. Readers usually express their impression by making a paraphrase to reaction terms. Thus, there are small amount of reactions that contain q_I directly.

The accuracy of the results depended on the query and the task. For instance, our system worked well on the query "aha, word-origin". The participants want to find documents about the etymology of terms that make the participants surprised for this query. (For instance, the term "salad" comes from the term "salt". Readers may feel aha after know this etymology.) In this case, the baseline 1 found many diaries and blogs. Some of those pages described in the way that "I felt AHA when I know the word-origin of ..." but did not contain the details of it. In other irrelevant pages, they use "aha" as a catch-phrase of page. In contrast, our system found a number of dictionary pages about the word-origin of terms. These pages were written in formal language without unnecessary sentences. In other words, they do not contain any impressions terms. The dataset contain 16 relevant pages that were evaluated by participant as impression relevant. Only two pages of them contained the term "aha" inside their main text. Some reactions of these pages contain the term "aha" directly, but the ratio of them is not so high. From this perspective, the baseline 2 could find such kind of pages, but still returned only a small part of pages compared with proposed method. Our method was able to expand the terms in reactions accurately. The expansion of the impression query brings higher recall without decreasing precision.

Another good result for our proposed method was for the task of "want-to-drink, cocktail", while baseline systems showed poor performance in this task. This result sheds light on the search problem on advertisement messages in Web pages. That is, "want-to-drink" is used as advertising statements very often. Thus some pages about new merchandises included this term, but participants did not agree with them. A similar pattern was seen in "want-to-go, restaurant Kyoto" task.

However, our system did not always show good results. In the case of the query "gammy, junior-high memory", the baseline system found blogs that organize comment in online forum sites. These blogs had comment columns for readers to leave their impressions. The baseline 1, *i.e.*, traditional AND search was useful enough to find blog articles from these

impression terms. In Twitter reactions, the term "gammy" does not appear frequently. Our system computed $s_w(w_j, q_{\mathrm{I}})$ on the basis of a few reactions. There tends to be fewer dismissive reactions than affirmative reactions in Twitter. The cause of this tendency is speculated to be that negative image words are avoided as conversation topics. Twitter users may desire smooth communications; thus they might simply ignore Web pages recommended by their friends instead of taking risk of issuing argument. Therefore, the proposed system could not estimate reaction terms from q_{I} accurately when the q_{I} is negative or dismissive term, since there was much less seed information for making a decision. In the "bad-looking, cat photo" task and in the "awful, dog pet photo" tasks, our system worked not very well for the same reason.

3.5 Summary

We proposed a novel Web search approach, "Search-by-Reaction" that exploits readers' impressions and evaluations. Search-by-Reaction makes it possible to query by users' subjective impression terms even if the target document does not contain these terms literally. Reactions to Web contents of people posting on Twitter are used as annotations of those Web pages. Reactions usually do not contain impression terms despite the fact that they express the same impressions in other words. Our system constructed the connection between impression terms and reaction terms.

We conducted experiments to show the effectiveness of our Search-by-Reaction model. Participants created a correct answer set as the gold standard, and compared our system with a method used in conventional search engines (AND search only focusing on terms in the main text) and a simplified version of our Search-by-Reaction system. Our system retrieved more relevant pages than did baseline systems, and the precision of our method was higher.

Our future work will be continuing to improve the performance of our system. Admittedly, there are several limitations of our system. First, our system are only able to retrieve Web pages that get one or more reactions in Twitter. So far, our system can find only a small number of Web pages. Therefore, it might be the future work to be predicting the reactions of Web pages by exploring the contents of pages. Moreover, we will concentrate on personal information such as the profiles of searchers and annotators for personalization so as to improve accuracy. The system can attach high weight to the reaction by the personwho is similar to the searcher. We will collect the reactions from more diverse Web sites, not only Twitter. By leveraging reactions posted on multiple Web communication sources can help to overcome the ranking biases of the annotators.

CHAPTER 4

Diversity-Based Credibility Analysis

4.1 Introduction

The rapid growth of the World Wide Web and Internet shopping services has enabled users to select from a huge number of commercial products on the Internet. Thus, the importance of user review data has increased, as it provides opinions and impressions that help users choose a quality item. There are many reviews for a variety of items on the Web, some of which are authored by professionals and others that are authored by non-professionals. Since professional reviews are available only for a limited number of items, even non-professional reviews are also useful for users to help making a decision.

However, there is a problem of *credibility* in utilizing reviews of general users. Since user reviews can be posted by any kinds of users including experts, novices, and even spammers, each review and aggregation of reviews can be biased and different from what the general public feels. Even users familiar with a particular domain cannot always produce a widely acceptable review, as they can be highly accustomed and accordingly biased to the domain. For example, users who have watched many Science Fiction (SF) movies might be likely to give a lower score to a SF movie than ordinary users, since they know more high-quality SF movies and use them as the basis for evaluating the other SF movies.

In this chapter, we focus particularly on the credibility of reviewers, where the credibility of reviewers is defined as the ability of precisely estimating the item quality. This ability is defined for a single reviewer, as well as a group of reviewers where the quality of items is

estimated by aggregated reviews (*e.g.*, the mean of their review scores). Thus, two problems regarding credibility are addressed in this chapter: 1) estimating the credibility of a single reviewer, and 2) estimating the credibility of a group of reviewer.

We tackle the first problem to discover *experts* based on their review experience approximated by the number of reviews, as well as diversity of reviewed items. Although the credibility of a reviewer possibly correlates to the number of reviews that he has posted, many reviews do not always guarantee high credibility of a reviewer. As we discussed earlier, users who have reviewed only a specific category of items might post highly biased reviews. Therefore, we also consider the diversity of reviewed items to accurately estimate the reviewer credibility, assuming that a reviewer who has reviewed in diverse categories has higher credibility. For example, we expect that users who reviewed a wide variety of movies have a higher ability to evaluate the quality of movies than those who reviewed only SF movies.

We tackle the second problem to precisely estimate the quality of items by aggregating reviews of a reviewer group. Even if the credibility of individuals is low, it is possible to achieve high credibility when their reviews are aggregated. This phenomenon is known as *the wisdom of crowds* [57], in which one of the key criteria to obtain quality results is diversity of opinions. Thus, our proposed method to estimate the credibility of a reviewer group stands on diversity of reviewers, with an assumption that a group of more diverse reviewers has higher credibility.

To verify the two assumptions mentioned above, we conducted experiments with a movie review dataset. The credibility of reviewers was measured by the similarity between their review score and a *true* score, which was approximated by the score given by a well-known professional reviewer. Our experimental results showed that the diversity of reviewed items and reviewers in a group was effective to estimate the credibility of a reviewer and a group of reviewers, respectively. Therefore, yes, the diversity does improve the credibility of user review data.

4.2 Methods

This section introduces methods to estimate the credibility of a reviewer and a reviewer group based on diversity measures. Our methods are designed to be applicable to a wide variety of user review data such as movies, hotels, books, restaurants, *etc.*

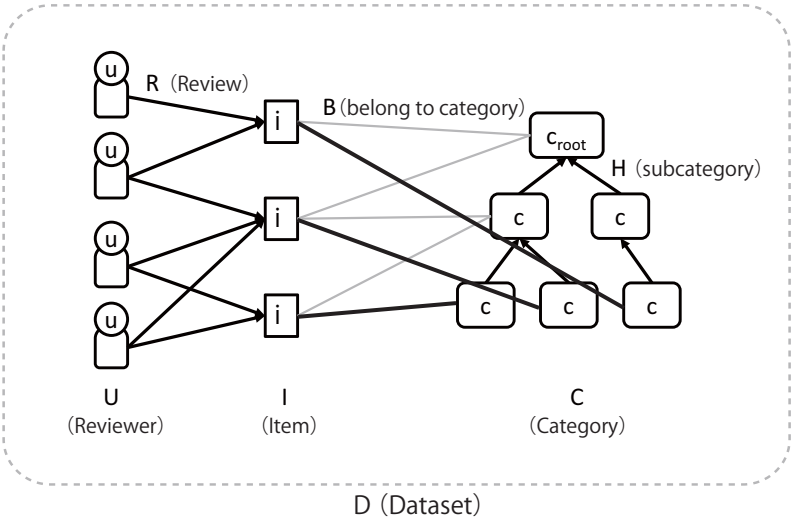

Figure 4.1: Review data structure.

4.2.1 User Review Data

User review data can be modeled by a tripartite graph with a category hierarchy. The tripartite graph consists of reviewers, items, categories, as well as reviewer-item and item-category edges. Figure 4.1 shows the structure of the general review dataset. The category hierarchy is a set of category-category edges. More specifically, user review data D is defined as follows:

$$D = (U, R, I, B, C, H), \tag{4.1}$$

where U is a set of reviewers, I is a set of items, C is a set of categories. A set of edges $R \subset U \times I$ represents reviews of reviewers for items, *e.g.*, $(u, i) \in R$ indicates that reviewer u reviewed item i. A set of edges $B \subset I \times C$ represents categories of items, *e.g.*, $(i, c) \in B$ indicates that item i belongs to category c. A set of edges $H \subset C \times C$ represents *is-a* relationships between pairs of categories, *e.g.*, $(c_j, c_k) \in H$ indicates that category c_j is a sub-category of category c_k.

Category tree $T = (C, H)$ is a rooted tree whose root is $c_{\text{root}} \in C$. Children of c_{root}, *e.g.*, $M = \{c \mid c \in C \wedge (c, c_{\text{root}}) \in H\}$, are called *main categories* and distinguished from the other categories.

Some variables used in our proposed methods are defined below. The number of reviews given by user u is defined as follows:

$$n_u = |\{i \mid i \in I \wedge (u, i) \in R\}|. \tag{4.2}$$

The number of items that belong to category c is defined as follows:

$$n_c = |\{i \mid i \in I \wedge (i, c) \in B\}|. \tag{4.3}$$

Finally, we define the number of items that belong to category c and have been reviewed by user u as follows:

$$n_{u,c} = |\{i \mid i \in I \wedge (u, i) \in R \wedge (i, c) \in B\}|. \tag{4.4}$$

4.2.2 Estimating the Credibility of a Reviewer

The first problem we tackle is to estimate the credibility of each reviewer. Recall that the credibility is the ability of precisely estimating the quality of items. Our method is based on the assumption that a reviewer who reviewed many and diverse items has high credibility. The reason why we came up with this assumption is explained as follows. Suppose that there are two reviewers: one reviewed ten movies, while another reviewed 100 movies. According to the assumption, the latter reviewer is more credible, as his expertise is expected to be higher than the former reviewer. Then suppose that there are another pair of reviewers: one reviewed 100 SF movies, while another reviewed 100 a wide variety of movies. We assume that the latter is more credible since his review is expected to be unbiased compared to the former reviewer.

The following formula is derived if we follow the assumption on the credibility of individual reviewer:

$$\text{Credibility}(u) = \alpha n_u \text{Div}(u), \tag{4.5}$$

where α is a parameter, n_u is the number of items reviewed by user u, and $\text{Div}(u)$ is the diversity of items reviewed by user u. We then model the diversity of reviewed items based on the idea of Shannon-Wiener index [46], which measures the diversity by the entropy over species. Regarding main categories as species in our case, Shannon-Wiener index is defined as follows:

$$S(u) = -\sum_{c \in M} p_u(c) \log p_u(c), \tag{4.6}$$

where $p_u(c)$ is the probability that user u reviews an item that belongs to category c. This probability can be estimated by the number of items of category c reviewed by user u divided by the number of items reviewed by user u: $p_u(c) = n_{u,c}/n_u$.

One of the problems of Shannon-Wiener index is that it is agnostic about the prior category distribution. Suppose that there are 10 horror and 100 SF movies. Although the maximum entropy is achieved by reviewing ten horror and ten SF movies, this reviewer is considered as biased to horror movies, as he reviewed all the horror movies despite the small number of horror ones. Therefore, we slightly modify Shannon-Wiener index by taking into account the prior category distribution. More specifically, we measure the diversity by the difference of the category distribution of a reviewer from the prior category distribution, *e.g.*, Kullback-Leibler divergence of the two distributions. Letting $p(c)$ be the prior category probability, Kullback-Leibler divergence is defined as follows:

$$\mathrm{KL}(u) = -\sum_{c \in M} p_u(c) \log \frac{p_u(c)}{p(c)}, \tag{4.7}$$

where $p(c)$ is the number of items of category c divided by the number of items: $p(c) = n_c/|I|$.

Finally, we define the diversity of a reviewer as follows:

$$\mathrm{Div}(u) = \exp(-\mathrm{KL}(u)). \tag{4.8}$$

Note that the exponential function is not essential, but is applied to make the diversity function $\mathrm{Div}(u)$ positively correlate to the diversity. This diversity function becomes larger when the category distribution of a reviewer and prior category distribution are closer. Thus, a reviewer who has evenly reviewed items is considered as credible, as he is considered as unbiased to any category.

4.2.3 Estimating the Credibility of a Group of Reviewers

The second problem we address is to estimate the credibility of a group of reviewers. Even if the credibility of individual reviewers is not so high, the credibility of a group of reviewers can be high when their reviews are aggregated. For example, the average review score of a group can be close to true quality of items, even if no individual reviewer can precisely estimate the quality.

According to the previous studies on collective intelligence [57], the diversity of members in a group is an important factor to obtain a high-quality result from the group by means of aggregation. For example, there are two groups: one includes ten SF maniacs, while another includes five SF and five horror maniacs. Given an item to each group, the average review

score given by the former group might be more biased than the latter, as the aggregated score may reflect only a specific preference of the homogeneous group.

Therefore, we propose methods to estimate the credibility of a reviewer group based on the diversity of the reviewers. Our assumption for this problem is that a group of many and diverse reviewers has high credibility. As the diversity can be measured by three types of aspects, namely, balance, variety, and disparity [56], we propose two diversity measures that take into account different aspects, *e.g.*, entropy-based and variance-based diversity measures.

The entropy-based diversity measure is similar to the one we used in estimating the credibility of individual reviewers, and takes into account the balance and variety of reviewers*. A high entropy-based diversity measure indicates that there are more types of reviewers in a group and the distribution of reviewers is balanced across the types. On the other hand, the variance-based diversity measure reflects the disparity aspect of diversity, and becomes high if reviewers in a group are dissimilar each other.

To compute the two diversity measures briefly explained above, it is necessary to model the similarity between reviewers in some way. To this end, we opted to characterize reviewers by using their expertise estimated by their reviews, with an assumption that a reviewer who has reviewed diverse items in a category has high expertise in the category. For instance, a reviewer who have watched and reviewed all of space opera, cyberpunk, and science fantasy movies is expected to have more knowledge in the SF category than one who have reviewed only space opera movies.

In a similar way to the diversity computation for a single reviewer, the expertise of user u in main category c is measured by Kullback-Leibler divergence of the sub-category distribution of a reviewer and prior sub-category distribution:

$$\mathrm{KL}_{\mathrm{sub}}(u, c) = -\sum_{s \in \mathrm{Sub}(c)} p_u(s|c) \log \frac{p_u(s|c)}{p(s|c)}, \tag{4.9}$$

where $\mathrm{Sub}(c)$ is a set of sub-categories of main category c (*e.g.*, $\mathrm{Sub}(c) = \{s \mid s \in C \wedge (s, c) \in H\}$), $p_u(s|c)$ is the probability that user u reviews an item of category s conditioned by category c ($p_u(s|c) = p_u(s)/p_u(c)$), and $p(s|c)$ is the prior probability of category c conditioned by category c ($p(s|c) = p(s)/p(c)$).

As the Kullback-Leibler divergence negatively correlates to the expertise in a main category, we apply an exponential function in the same way as the diversity computation

*Balance and variety are simultaneously measured since they are not divisible in many cases.

for a single reviewer, and define the expertise of user u in main category c as follows:

$$e_{u,c} = \exp(-\mathrm{KL}_{\mathrm{sub}}(u, c)). \tag{4.10}$$

Below, we explain the two diversity measures in the details.

Entropy-based Diversity Measure

The entropy-based diversity measure is the entropy of the expertise distribution of a group as a whole with consideration of the prior expertise distribution. We first model the expertise of group $G \subset U$ in category c by aggregating the expertise of reviewers in the group:

$$e_{G,c} = \frac{1}{|G|} \sum_{u \in G} e_{u,c}. \tag{4.11}$$

We then model the *prior* expertise in category c:

$$e_c = \frac{1}{|U|} \sum_{u \in U} e_{u,c}. \tag{4.12}$$

The prior expertise can be interpreted as the average expertise in all the reviewers. Although these expertise scores do not represent a probability, we could normalize the expertise scores to treat them as probabilities:

$$p^e_G(c) = \frac{1}{|G|} \sum_{u \in G} e_{u,c}, \tag{4.13}$$

$$p^e_c = \frac{1}{|U|} \sum_{u \in U} e_{u,c}. \tag{4.14}$$

Kullback-Leibler divergence of the expertise distribution of a reviewer group and the prior expertise distribution is defined as follows:

$$\mathrm{KL}^e(G) = -\sum_{c \in M} p^e_G(c) \log \frac{p^e_G(c)}{p^e(c)}. \tag{4.15}$$

This divergence represents the closeness between the expertise of a group and prior expertise, and becomes smaller if the group expertise is more evenly distributed against the prior expertise.

Entropy-based diversity measure EDiv is then defined as follows:

$$\mathrm{EDiv}(G) = \exp(-\mathrm{KL}^e(G)). \tag{4.16}$$

Note that the exponential function is not essential again.

The entropy-based diversity measure increases as the expertise of a group as a whole is evenly distributed in each category. Note that this measure does not take into account the diversity of each reviewer in a group, and becomes high in both of the following cases: 1) all the reviewers in the group have balanced expertise in each category, and 2) the expertise distribution of the group is close to the prior expertise distribution, even though the expertise distribution of each reviewer is far from the prior expertise distribution.

Variance-based Diversity Measure

As computing the variance-based diversity measure requires the dissimilarity between reviewers, we first map reviewers on a $|M|$-dimensional space, where each dimension represent the expertise in a main category. A vector of reviewer u is denoted by $\mathbf{v}_u$ and defined as follows:

$$\mathbf{v}_u = (e_{u,c_1}, e_{u,c_2}, \ldots, e_{u,c_{|C|}}), \tag{4.17}$$

where $e_{u,c}$ is the expertise of reviewer u in category c.

Variance-based diversity measure VDiv, which is the average dissimilarity between individual reviewers and the mean of the reviewers in the group, is defined as follows:

$$\mathrm{VDiv}(G) = \frac{1}{|G|} \sum_{u \in G} \|\mathbf{v}_u - \bar{\mathbf{v}}_G\|, \tag{4.18}$$

where $\bar{\mathbf{v}}_G$ is the mean of reviewer vectors of group G, *i.e.*, $\bar{\mathbf{v}}_G = \frac{1}{|G|} \sum_{u \in G} \mathbf{v}_u$.

In summary, we proposed diversity measures to estimate the credibility of reviewers as an individual and as a group. A variant of Shannon-Wiener index was proposed to measure the diversity for both of the cases, and a variance-based diversity measure was used only for a reviewer group. Note that the entropy-based and variance-based diversity measures correlate to some extent, but behave differently in some cases. For example, the entropy-based diversity measure becomes high if reviewers in a group have similar expertise in a wide variety of categories, whereas the variance-based diversity measure does not. In the next section, we demonstrate the correlation between the credibility and diversity measured by the proposed methods.

4.3 Experiments

To clarify the effectiveness of our diversity measures for estimating the credibility of reviewers, we conducted experiments by using movie review data taken from Yahoo! Movies. Through the experiments, we tested the validity of the two assumptions: 1) a reviewer who has reviewed many and diverse items has high credibility, and 2) a group of reviewers is credible if the group consists of many and diverse reviewers.

4.3.1 Dataset

The movie review data was taken from Yahoo! Movies[†], which is one of the biggest movie communities in Japan. We collected 27,516 movies and 158,385 reviewers. There are 1,124,555 reviews and 38 categories in this dataset.

Since some real review data including ours do not contain explicit hierarchy information in categories, we applied a heuristic method to construct a hierarchy. Our method first extracted existing categories as main categories (*e.g.*, 38 categories in our data), and then generated sub-categories by combining any pair of co-occurring main categories. More precisely, letting M be a set of main categories, we define a set of sub-categories as $S = \{c_j \oplus c_k \mid i \in I \wedge (i, c_j) \in B \wedge (i, c_k) \in B\}$, where $\oplus$ is an operator to concatenate two category names. We let the resultant set of sub-categories belong to main categories from which the sub-categories were generated, *e.g.*, edges (c, c_j) and (c, c_k) were added to H for $c = c_j \oplus c_k$. For example, "Star Wars" belongs to two main categories *SF* and *adventure*. We created a sub-category *SF* - *adventure* and let it belong to *SF* and *adventure*. Finally, a set of categories is defined as $C = M \cup S$.

Note that we created a special sub-category indicating that a movie belongs to only a main category and does not belong to any sub-category. Given a movie belonging only to main category c, we added subcategory $c' = c \oplus c$ to the entire category set, and edge (c', c) to H. This special type of sub-categories was added because movies without any sub-category are not taken into account in the expertise estimation. For instance, the movie "Blade Runner" belongs only to *SF* category. This movie was assigned to a *SF* - *SF* sub-category.

Tables 4.1 and 4.2 show the detailed statistics of reviewers and movies in our dataset, from which we can find many reviewers who posted a review only once, and movies with a few reviews.

4.3.2 Evaluating the Credibility of a Reviewer

The first assumption is that a reviewer who has reviewed many and diverse items has high credibility. To test this assumption, we compared the correlation between the credibility and following measures: quantity (n_u in Equation 4.2), diversity (Div(u) in Equation 4.10), and both diversity and quantity (Credibility(u) in Equation 4.5 ($\alpha = 1$)).

Before testing the first assumption, we start with illustrating the characteristics of these three measures. Figure 4.2 shows how well the three measure distinguish expert reviewers

[†]http://movies.yahoo.co.jp/

# of reviewers	
Reviewed only 1 movie	140,180
Reviewed less than 10 movies	204,178
Reviewed 1,000+ movies	39
Reviewed 2,000+ movies	7
Total	220,471

# of movies	
Reviewed by only 1 Reviewer	6,326
Reviewed by 10+ Reviewers	8,877
Reviewed by 1,000+ Reviewers	158
Reviewed by 2,000+ Reviewers	29
Total	27,514

# of reviews per reviewer	
Arithmetic mean	6.35
Mode	1
Median	1
Max	5,301

Table 4.1: Statistics of reviewers.

# of reviewers per movie	
Arithmetic mean	40.82
Mode	1
Median	4
Max	6,304

Table 4.2: Statistics of movies.

from the others, where the horizontal axis represents the value of each measure, and the vertical axis represents the entropy of review scores. Each point in the figures represents the value of a measure and review score entropy of a reviewer. According to McAuley and Leskovec's work, experienced reviewers have a higher review score entropy, while novice reviewers cannot take full advantage of the range of scores, and are likely to evaluate items in a narrow and biased manner. For example, novice reviewers may use only three or four even if they are asked to evaluate movies at a five-point scale. Thus, the review score entropy can be a good indicator of experts.

In the ideal case, points in the figures should converge towards the upper right corner: some novice reviewers gave a wide or a narrow range of scores, while the most expert reviewers gave a wide range of scores. It can be seen from Figure 4.2 that both of the quantity and diversity can distinguish experts (reviewers with high review entropy) from the others. The diversity measure shows a slightly better discriminative power as reviewers broadly spread along the horizontal axis.

To test the first assumption, it is necessary to obtain *true* quality of each item. Since it is hard to get exact true quality score, we approximated it by the score given by a well-known professional reviewer. We extensively compared reviewers who rated many and diverse movies, and carefully selected one who gives a widely acceptable score. Finally, we decided to use reviews authored by Yuichi Maeda, and manually collected his reviews from

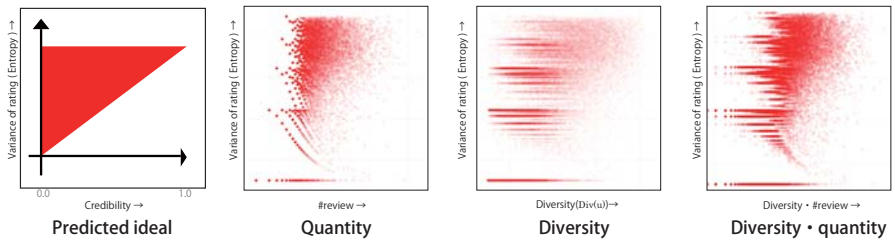

Figure 4.2: Quantity, diversity, and their combination vs. review score entropy.

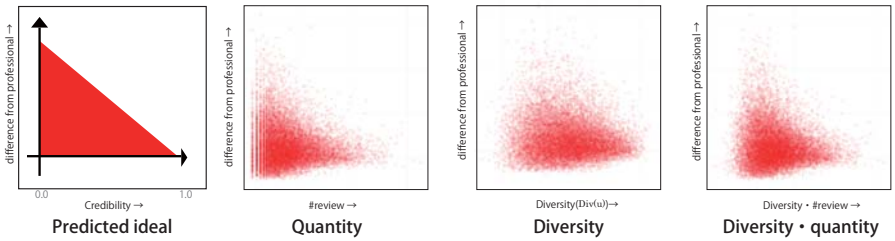

Figure 4.3: Quantity, diversity, and their combination vs. RSS to professional scores.

his Web site[‡]. He is a Japanese professional critic and movie journalist who has written 1,832 reviews since 2003 to 2014 on his site. We found 1,689 movies included in both of his and our review data. As the range of his review scores was different from ours, we converted them to a five-point scale and used the scores as true quality of items.

The credibility of a reviewer was estimated by the residual sum of squares (RSS) between his score and a score of reviewer u:

$$\mathrm{RSS}(u) = \frac{1}{|I_u \cap P|} \sum_{i \in I_u \cap P} (\mathrm{score}(u,i) - \mathrm{score}_{\mathrm{pro}}(i))^2, \tag{4.19}$$

where P is a set of movies reviewed by the professional, I_u is a set of movies reviewed by user u ($I_u = \{i \mid i \in I \wedge (u,i) \in R\}$), $\mathrm{score}(u,i)$ is a review score of u for movie i, and $\mathrm{score}_{\mathrm{pro}}(i)$ is a review score of the professional for movie i.

Figure 4.3 demonstrates that a reviewer becomes more similar in rating to the professional reviewer if the reviewer has reviewed more and more diverse movies.

[‡]http://movie.maeda-y.com/

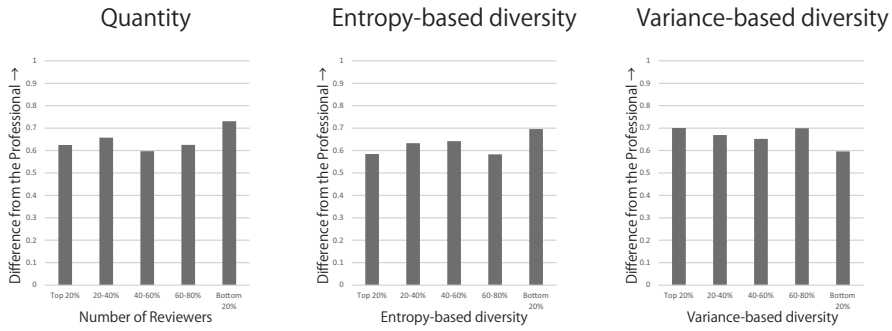

Figure 4.4: Quantity, entropy-based, and variance-based diversity measure vs. RSS to professional scores (for all the groups).

4.3.3 Evaluating the Credibility of a Group of Reviewers

To test the second assumption regarding the credibility of a group of reviewers, we compared following measures: quantity ($|\{u \mid u \in U \wedge (u,i) \in R\}|$ for item i), entropy-based diversity measure (EDiv(G) in Equation 4.16), and variance-based diversity measure (VDiv(G) in Equation 4.18). The absolute error between the score of the professional and the average score of group G for item i is defined as follows:

$$\mathrm{AE}(G,i) = \frac{1}{|G|}\sum_{u\in G} |\mathrm{score}(u,i) - \mathrm{score}_{\mathrm{pro}}(i)|\,. \tag{4.20}$$

If our second assumption is probable, the absolute error from large and diverse groups is smaller than that of smaller and/or more homogeneous groups.

Figure 4.4 shows the average absolute error of groups in each *bin*. We sorted all the groups based on one of the three measures, and categorized them into five bins based on the order of groups. For example, the leftmost bin of each figure includes groups ranked within top 20% when they are sorted in descending order of each measure. Thus, the left bins of each graph contain reviewer groups that are estimated as more credible, whereas the right bins contain reviewer groups that are estimated as less credible.

In the ideal case, the bars would slant upward to the right: the absolute error to the professional should become bigger for smaller or more homogeneous groups, while the error should be smaller for bigger or more diverse groups. The graph of a group whose members are many or diverse is close to it. The bars of the quantity and entropy-based diversity measure show slightly similar trends to the ideal case, though they are not conclusive. The

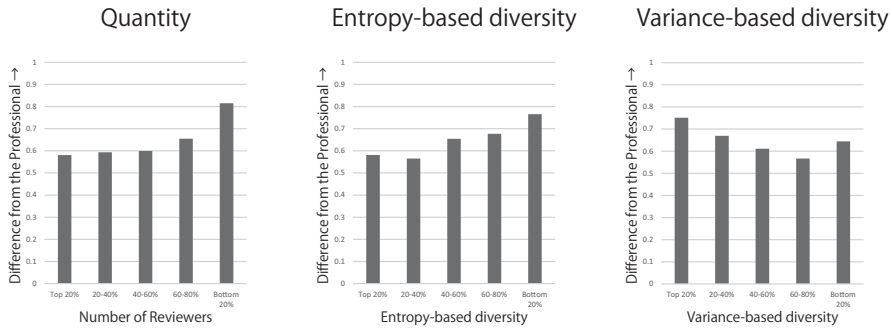

Figure 4.5: Quantity, entropy-based, and variance-based diversity measure vs. RSS to professional scores (for groups with less than 100 reviewers).

graph of the variance-based diversity measure does not show a trend similar to the ideal case. When we compare the leftmost bins, which contains the most diverse groups (top 20%), it can be seen that the entropy-based diversity measure outperforms the quantity-based measure in finding the most credible reviewers.

As we have observed from Figure 4.4, there is a notable absolute error difference between groups with different diversity. We hypothesized that the absolute error for the professional can be small enough if plenty of reviews were available for each movie, and investigated a case where a limited number of reviews are available. Figure 4.5 shows the average absolute error of groups with less than 100 reviews. In this case, the entropy-based diversity measure and the number of reviewers can more accurately estimate the credibility of reviewer groups.

4.4 Discussion

Our first experiment was successful in evaluating the credibility of a reviewer, supporting our hypothesis that reviewers who see diverse movies and reviewers who see many movies are reliable. We learned that these reviewers are characterized by a more homogeneous distribution of review scores and for a smaller difference in rating with professional reviewers.

The reason why the difference of opinion of amateurs and professionals does not converge to zero is the difference of average; the professional's average rating is 3.3, and amateurs' is 3.6. Professionals are sometimes forced to see and to rate unfavorite movies at the job. Amateur can choose their favorite movies to review. In such situations, the set of movies reviewed by a reviewer is often biased toward highly acclaimed or well personalized movies.

From the second experiment, we established that the entropy-based diversity and re-

viewer group size are good guideline to measure the credibility of a group. In contrast, Variance-based diversity does not work well.

The entropy-based diversity can measure the credibility of a group especially when the number of reviewer is lower than 100. It is interesting to note that, when the number of members is small, the diversity of members is important, but when it is large, this is not the case. Generally, when the size of a group is large enough, most of groups are reliable when it is likely that the credibility of the group is saturated, we do not need to consider the size and diversity of the group. Figure 4.6 shows the relationship between the effect of the entropy-based diversity and the size of a group. The horizontal axis lists the groups binned by size. Each bin contains same number of groups. Groups were classified into high-diversity groups and low-diversity groups by their median entropy-based diversity. The vertical axis shows the average distance between the rating of the professional and the one of the group. When the number of reviewers is less or equal to 440, a high diversity of reviewers minimized the score difference with the professional review. This fact supports our proposition. On the contrary, in cases where the number of reviewers exceeds 440, the diversity of reviewers did not affect the score difference. Naturally, a larger group will be more credible because of the law of large numbers. The accuracy of the average score, however, trended down for the cluster of movies that assumed 119 to 182 reviews. This can be attributed to two possible causes. The movie reviewed by many reviewers is a popular movie, who tend to attract an audience of persons unfamiliar with movies. Their opinions are not very credible as evidenced by professional reviewers often shooting down popular movies. It reflects a characteristic of the review dataset; online user review are not implicit data, but intentional data.

The variance-based diversity does not work well, regardless of the group size. One reason could be a biased group (*e.g.*, a community of specialists) providing a correct opinion. Another cause could be generalists. They are similar to each other. A group that consists of non-diverse generalists can rate movies accurately.

4.5 Summary

In this chapter, we proposed a method to estimate credibility of individuals and reviewer groups. We proposed two simple assumptions: a reviewer who has reviewed many and diverse items has a high credibility, and a group of reviewers is credible if the group consists of many and diverse reviewers. We modeled a general user review structure with a category tree and proposed diversity-based measurement calculations. Through experiments using a real dataset of movie reviews, the effectiveness of the assumption 1 was confirmed; a

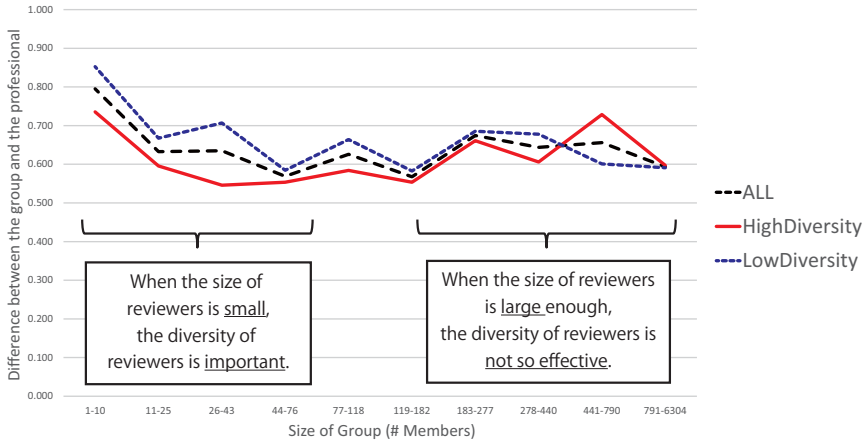

Figure 4.6: Effectiveness of the entropy-based diversity.

reviewer, who reviews many and diverse movies has a high credibility. The effectiveness assumption 2 was partially confirmed; when the number of members is small, the entropy-based diversity is a good indicator to measure the credibility of a group.

CHAPTER 5

Diversity-Based HITS

5.1 Introduction

As the Web continues to rapidly expand, both Internet users and the purposes of Web documents are becoming more extensive and diverse. Users of the Web are not only computer specialists but also ordinary people, and the content on the Web has accordingly become broader, including not just informative documents but also many sub-products of communication, personal diaries, and so on. This has made Web usage increasingly more complex.

Many major Web search engines use link analysis algorithms such as HITS and PageRank to rank Web documents. For example, Google uses PageRank, which calculates popularity scores. These methods focus on the number of linking documents of each page. The popularity score is determined by recursive calculation using the simple hypothesis that pages linked by many popular pages are popular pages. The HITS algorithm uses the number of linking and linked pages to calculate the Hubs and the Authority. However, these methods focus only on the number of linking pages and are therefore unsuitable for coping with the demands of complex usages of the Web.

For example, when a novice user wants to know what a "compiler" is and inputs the query "compiler" to a search engine, the result contains many different kinds of pages, such as dictionary pages (*e.g.*, thesaurus), encyclopedia pages (*e.g.*, Wikipedia), introductory articles, commercial sites about specialized compilers, academic articles, and so on. All of them are popular documents and contain the term "compiler". However, the most popular documents are not necessarily useful for all users. In this case, dictionary, encyclopedia

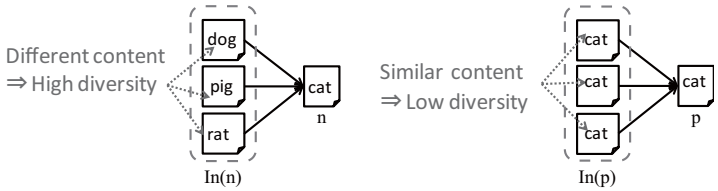

Figure 5.1: Referrer diversity.

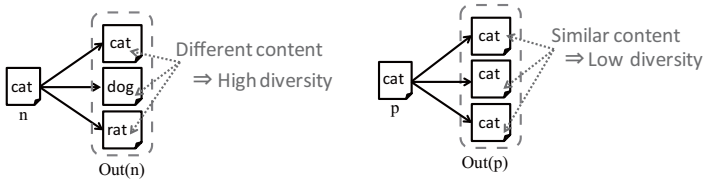

Figure 5.2: Referral diversity.

pages and introductory articles would be useful for novice and general users, but other pages would be more useful for a limited number of specialists. The problem here is that the existing link analytic methods score every document based on how many documents link to them instead of checking how it was linked. The number of linking documents expresses how popular the document is, but does not express why it is popular.

In this chapter, we propose a new link analysis algorithm that considers not only the number of linking pages but also the diversity of linking pages and linked pages in order to consider the reason why the page is popular.

Figure 5.1 shows an example of different ways a page is linked. There are two pages about cats, both of which are linked by three pages. Page n is the first page about cats, which is referred by documents on three different topics: a page about dogs, a page about pigs, and a page about rats. Page p is also linked by three pages, but all of them are about cats. Since each linking page has the same popularity score, PageRank gives the same popularity score to page p and page n because both of them are linked by the same number of pages. However, we suspect there is a big difference in the reason behind their popularity. Page n is linked by diverse pages. The authors of pages that link other pages may just be readers of the latter, who create the link after viewing such pages and finding them interesting. When the topic in the document reflects the interest of the author, a page linked by diverse pages must have a wide readership. An article with a diverse readership is assumed to contain information that can be interesting for many people:

general information, universal information, and so on. Such information is useful not only for specialists but also for novice users. In contrast, page p has a biased, non-diverse readership, making it suitable for certain specialists or members of a specific community. The same can be said of linking sites and navigating sites (see Figure 5.2). In this case, the former linking page may have been made by a user who has a broad outlook and the latter by one whose range of interest is narrow. These examples demonstrate that by considering how the linking pages and linked pages are diverse, the search algorithm can create a ranking that depends on the reason behind the popularity.

We define diversity as the dispersion of a set of pages. When each page in the set has a different topic, the diversity score is high, and when all pages are similar, the diversity score is low. We expand the HITS algorithm, which is the standard existing link analysis algorithm, with two types of diversity score: *referrer diversity* and *referral diversity*. Referrer diversity means how pages that link the page are diverse and referral diversity means how pages linked by the page are diverse. The conventional HITS algorithm calculates Hubs, which means how the page links many good authorities, and Authorities, which means how the page is linked by many good hubs. We expand the concept of Hubs and Authorities with diversity by two simple hypotheses:

- The page linking diverse Authorities is a valuable Hub
  (This Hub can be created by a well-informed generalist who has a wide range of interests.)
- The page linked by diverse Hubs is a valuable Authority
  (It must be useful not only in a specific field.)

5.2 Methods

In this section, we explain our link analysis method based on diversity in detail. The proposed method is composed of two parts. The first is calculating the diversity of the set of pages. For this part, we propose a method to quantify how each page in the set is different from the others. This quantification method creates a feature vector of each page with Latent Dirichlet Allocation (LDA) and then uses the sum of the distance between the centroid and each page in the set. The second part is expanding the HITS algorithm by using the diversity of the referrer and referral documents of each document. The method calculates the diversity of document links to the document as the *referrer diversity* and of the document linked by the document as the *referral diversity*. We set these two diversity scores in the HITS algorithm as the weight of the edge.

The purpose of the proposed method is to find pages that are useful for everyone: not only specialists but novice users as well. We assume a simple HITS-based hypothesis: the Hub page that links to diverse Authorities and the Authority page that is linked by diverse Hubs must feature a wide readership and widespread interest and therefore be of general interest to everyone.

5.2.1 Determining Diversity

To calculate diversity, each document has to be expressed as a feature vector. In our method, we take topics from the main text of the document and define the diversity as how different the topics of a page set are. The simplest way to express the document as the vector is just counting all the terms in the main text, but when the number of documents increases, the number of vector dimensions explodes. We used LDA to compress the dimension. Each term in the dataset is assumed to belong to one topic of i types of topics based on the topic model, where i is the given number of dimensions. LDA classifies terms into the topics to which they belong. The frequency of topic occurrence in each document is used as the feature vector of the document. Each document can be expressed by an i-dimensional vector. The length of documents in the dataset is not constant, so the feature vector has to be normalized.

Note that to create a feature vector, we can use other information in addition to the topic of the main text, such as the degree of confirmation or denial, the stance of the author, sentiments, and so on [42]. If another kind of diversity is needed, the algorithm can deal with it by switching function just as well based on a differently constructed feature vector.

We propose a diversity function $d(P)$ to calculate the diversity of the document set P. This function takes a high value when each document in P has a different topic and a low value when all documents have a similar topic.

Every document is defined as $n = (n_1, n_2, n_3, n_4, \ldots, n_k)$, that is, documents in the dataset are explained as a k-dimension feature vector based on the frequency of the topic found in their main text. The diversity, which means how the documents in the set P are diverse, is defined as

$$d(P) = \frac{1}{|P|} \sum_{p \in P} dist(p, mean(P)), \tag{5.1}$$

where $mean(P)$ is the arithmetic mean of the set of documents P and $dist(a, b)$ is the Euclidean distance between vectors a and b. The arithmetic mean $mean(P)$ is

$$mean(P) = \frac{1}{|P|} \sum_{p \in P} p, \tag{5.2}$$

and distance $dist(a, b)$ is

$$dist(a, b) = \sqrt{\sum_{i=1}^{k}(a_i - b_i)^2}, \tag{5.3}$$

where a and b are vectors expressed as $a = (a_1, a_2, a_3, \ldots, a_k)$, $b = (b_1, b_2, b_3, \ldots, b_k)$. The definition of $d(P)$ in our method is the normalized sum of the difference between each page in P and its mean. When the dispersion of the documents in P is high, this value become high. The value drops into $[0, \sqrt{1 - \frac{1}{k}}]$ when the norm of the feature vectors is normalized to one. The diversity score $d(P)$ gets its maximum value when all the articles in P have a different topic and gets its minimum value, zero, when all documents in P have the same content or the number of documents in P is less than two. We call $d(In(n))$ the *referrer diversity* of n, which means how pages linking n are diverse, where $In(n)$ is the set of pages that link to n. Likewise, we call $d(Out(n))$ the *referral diversity* of n, which means how pages linked by page n are diverse, where $Out(n)$ is the set of pages linked by n. It is important to note that this calculus equation is not so novel. You can see the same equation in the K-means clustering as the value to minimize. It is used to express the cohesiveness of the cluster. When the dimension of the vector is one, this equation means the variance.

5.2.2 Diversity-Based HITS Algorithm

In this section, we explain the method to calculate the Hub score and the Authority score by using the diversity-based HITS algorithm considering referral diversity and referrer diversity.

First, we have to prepare the root set that is used for the result page of the given query. We also need a graph for the link analysis, so a base set was created as the sum set of the root set itself, linking document set and linked document set of the root set.

The original HITS algorithm defines Hubs and Authorities by mutual recursion, as

$$hub(p) = \sum_{q, p \to q} auth(q), \tag{5.4}$$

$$auth(p) = \sum_{q, q \to p} hub(q), \tag{5.5}$$

where both p and q is a page in the base set. It can be expressed as matrix calculation below:

$$h = \boldsymbol{A}a, \tag{5.6}$$

$$a = \boldsymbol{A}^T h, \tag{5.7}$$

where $\boldsymbol{A}$ is the adjacency matrix of the dataset and h and a are the vectors of the Hubs and Authorities, respectively.

$$\boldsymbol{A}_{ij} = \begin{cases} 1 & \text{if } i\text{-th page links } j\text{-th page,} \\ 0 & \text{otherwise.} \end{cases} \tag{5.8}$$

In this case, $\boldsymbol{A}_{ij}$ means the link between j-th page and i-th page in the dataset.

Our method modifies this HITS calculation by using diversity-based factors $d(In(p))$ and $d(Out(p))$ as

$$hub_{\mathrm{d}}(p) = d(Out(p)) \sum_{q, p \to q} auth_{\mathrm{d}}(q), \tag{5.9}$$

$$auth_{\mathrm{d}}(p) = d(In(p)) \sum_{q, q \to p} hub_{\mathrm{d}}(q), \tag{5.10}$$

where $In(p)$ is the set of pages that links page p, and $Out(p)$ is the set of pages linked by page p. Two scores: $hub_{\mathrm{d}}(p)$ and $auth_{\mathrm{d}}(p)$ mean diversity-based Hubs and diversity-based Authorities. We call $d(In(p))$ as *referrer diversity* of page p, and $d(Out(p))$ as *referral diversity* of page p. This formula supports the two diversity-based hypotheses above, that is, "The Hub that links diverse Authorities is a good Hub" and "The Authority that is linked by diverse Hubs is a good Authority."

We can replace the adjacency matrix $\boldsymbol{A}$ of the HITS algorithm. We propose two diversity-based adjacency matrixes. One is based on referral diversity and the other on referrer diversity. The expanded adjacency matrix taking referral diversity is as below:

$$\boldsymbol{N}_{ij} = \begin{cases} d(In(p_j)) & \text{if } i\text{-th page links } j\text{-th page,} \\ 0 & \text{otherwise,} \end{cases} \tag{5.11}$$

where $In(p_j)$ is the set of pages that links j-th page p_j. In this matrix, links to the document that are linked by many different documents are weighted highly and receive a higher score. In contrast, a higher score is not correlated with pages linked by many similar pages. The other diversity-based adjacency matrix, $\boldsymbol{O}$, which takes the referral diversity, is

$$\boldsymbol{O}_{ij} = \begin{cases} d(Out(p_i)) & \text{if } i\text{-th page links } j\text{-th page,} \\ 0 & \text{otherwise,} \end{cases} \tag{5.12}$$

where $Out(p_i)$ is the set of pages linked by page p_i. This matrix means that the weight of the link from the page linking diverse pages become high, and the weight of the link becomes low when the link is from a page linking similar pages. To consider referral diversity, $\boldsymbol{A}$ should be replaced with $\boldsymbol{O}$. $\boldsymbol{A}$ can be replaced with $\boldsymbol{N}$ to consider referrer diversity.

$$\boldsymbol{h}_{\mathrm{d}} = \boldsymbol{O}\boldsymbol{a}_{\mathrm{d}}, \tag{5.13}$$

$$\boldsymbol{a}_{\mathrm{d}} = \boldsymbol{N}^T\boldsymbol{h}_{\mathrm{d}}, \tag{5.14}$$

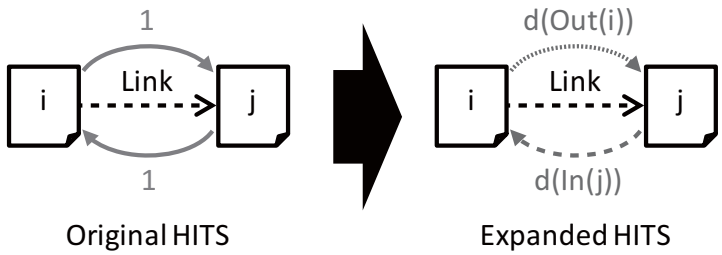

Figure 5.3: Asymmetric propagation value.

where $\boldsymbol{a}_{\mathrm{d}}$ and $\boldsymbol{h}_{\mathrm{d}}$ are the vectors of the diversity-based Hubs and diversity-based Authorities, respectively. In the original HITS algorithm, the coefficient of propagation is the same with the Hubs to Authorities propagation and the Authorities to Hubs propagation. The proposed method replaces $\boldsymbol{A}$ and $\boldsymbol{A}^T$ individually with diversity-based matrixes. It makes an asymmetric link weighted bipartite graph (see Figure 5.3). To calculate diversity-based HITS, it uses referral diversity score for back link propagation from Authority page to Hub page, and referrer diversity score for link propagation from Hub page to authority page.

The expanded HITS algorithm with a diversity-based adjacency matrix can be solved by the power method.

$$\boldsymbol{h}_{\mathrm{d}} = \boldsymbol{O}\boldsymbol{N}^T\boldsymbol{h}_{\mathrm{d}}, \tag{5.15}$$

$$\boldsymbol{a}_{\mathrm{d}} = \boldsymbol{N}^T\boldsymbol{O}\boldsymbol{a}_{\mathrm{d}}. \tag{5.16}$$

Vectors $\boldsymbol{a}_{\mathrm{d}}$ and $\boldsymbol{h}_{\mathrm{d}}$ converge to Authorities and Hubs when they are normalized by every phase. Authority-based ranking can be used to find documents and Hub-based ranking can be used to find good linking sites or navigating sites.

5.3 Experiments

To compare the methods explained in Section 5.2 with the original HITS algorithm and variant methods, we conducted an experimental Web ranking evaluation. The pages for the given query were sorted by the Authority score of each method as a search result ranking. Each of these rankings was evaluated by bucket-based evaluation. As stated previously, the aim of the proposed method is to enable both novice and expert users to find information that is useful. We used one participant who played a novice and classified documents sampled by each ranking into useful and useless documents after reading the main text.

5.3.1 Variant Methods

To clarify the effect of referral diversity and referrer diversity particularly, we have prepared two variant methods.

One is the referral diversity-based method. It replaces $\boldsymbol{A}$ with $\boldsymbol{O}$, and $\boldsymbol{A}^T$ is unchanged. It considers only referral diversity and not referrer diversity. This supports the hypothesis that the "The Hub that links diverse Authorities is a good Hub."

Another one is the referrer diversity-based method. It replaces $\boldsymbol{A}^T$ with $\boldsymbol{N}^T$, and $\boldsymbol{A}$ is unchanged. It considers only the referrer diversity. This supports the hypothesis that the "The Authority that is linked by diverse Hubs is a good Authority."

We compared them to proposed method as baseline methods.

5.3.2 Dataset

We used the ClueWeb09-JA dataset [60], which contains over 67 million pages with 400 million links between them. We prepared 26 queries, shown in Table 5.1. Nine queries of them were taken from author's search log. We chose nine queries from NTCIR-9 dataset, and eight queries from NTCIR-10 dataset*. The queries taken from NTCIR datasets were prepared for one-click task and intent task. The purpose of these tasks are to search definition of the terms. Thus, the most of these queries are general nouns. The top 1,000 pages on BM25 sorted ranking were extracted by each query as the root set. Pages linking to a page included in the root set and linked by pages in the root set were used as the base set. The root set was then sorted by each method. The page evaluated by the participant is a sample of the root set.

We compared four methods below.

- **Both** is the proposed method based on both diversity factors. Pages are ranked by the Authority score calculated by Equation 5.9. It uses the referrer diversity to calculate Authorities and the referral diversity to calculate Hubs.

- **Referrer** is the variant method based only on referrer diversity. It uses the referrer diversity factor on the propagation from Authorities to Hubs.

- **Referral** is the variant method based only on referral diversity. It uses the referral diversity factor on the propagation from Hubs to Authorities.

- **HITS** is the baseline method. It is the original HITS algorithm.

*http://research.nii.ac.jp/ntcir/index-ja.html

Each method is compared using the Authority-based score because in this experiment we want to find the document but not the linking page.

To compare rankings by these methods, we evaluated sample pages of each ranking. First, we split the ranking into five buckets: the top ten pages and pages ranked from 11 – 50, 51 – 200, 201 – 500, and 501 – 1,000. We took ten sample pages from each bucket. The sampling rate of each bucket was not constant: the upper part of the ranking was sampled in high density and the bottom part was sampled coarsely. All of the top ten pages were evaluated by one participant, but only 2 % of the bottom pages were evaluated. The total number of evaluated pages was 4,073 by 26 queries and four methods after removing duplicates. Two participants evaluated each document in terms of whether or not it was useful for a novice user. Sample pages were sorted randomly for every query. Each page in the dataset was shown as plain text.

We used GibbsLDA++[†] as an implementation of LDA. We classified the terms in the dataset into 100 topics. The LDA sampling was iterated 2,000 times.

5.3.3 Result

The results are shown in Table 5.2 and Fig. 5.4. The original HITS method ranked many correct pages in the middle of the ranking. All expanded methods were influenced by the original HITS method. In the ideal case, it is hoped that many correct pages appear in the top part of the ranking and that a small number of correct pages appear in the bottom. The **both** method, which uses both types of diversity, found more useful pages for novices in the top part of the ranking than the other methods. In the top bucket, there is significant difference in the precision of the method proposed and original HITS algorithm. The proposed method could find double correct pages on the top ten ranking. The p-value is 0.01. The difference of precision between two methods became small in the middle part of the ranking, and the original HITS algorithm found more correct answer in the bottom bucket. In other words, the method proposed could evaluate not suitable pages to low rank. The p-value in bottom bucket is 0.16, thus effectiveness of the proposed method is lower in the bottom part of ranking. The **referrer** method seems a little bit better than the original HITS method in the top part of ranking, but it ranks many correct pages in the bottom part. The **referral** method had almost the same accuracy as the original HITS method. Across the board, the ratio of correct pages in the dataset was just 11.7% through the all queries and all methods, that is, to find the pages suitable for novice users is hard task. The eight queries of original queries used in the experiment

[†]http://gibbslda.sourceforge.net/

Queries	Type
Inu-Yasha (comic's name)	NTCIR-9
GPS	NTCIR-9
The Secret Garden	NTCIR-9
AVP (Alien vs. Predator)	NTCIR-9
The Wall	NTCIR-9
Buzz Lightyear	NTCIR-9
Make haste slowly	NTCIR-9
Khufu	NTCIR-9
Parkinson's disease	NTCIR-9
Ika-Ten (TV show's name)	NTCIR-10
Electone	NTCIR-10
Ondulish (the Ondul language)	NTCIR-10
Sculpture	NTCIR-10
Hospitality	NTCIR-10
Ramsar Convention	NTCIR-10
Show humanity even to one's enemy	NTCIR-10
Exclusive economic zone	NTCIR-10
Postal service privatization	Original
France trip	Original
Sagrada Familia	Original
Fish called by different names in life stage	Original
Game theory	Original
Machine learning	Original
Compiler	Original
Neoliberalism	Original
Ohm's law	Original

Table 5.1: Queries.

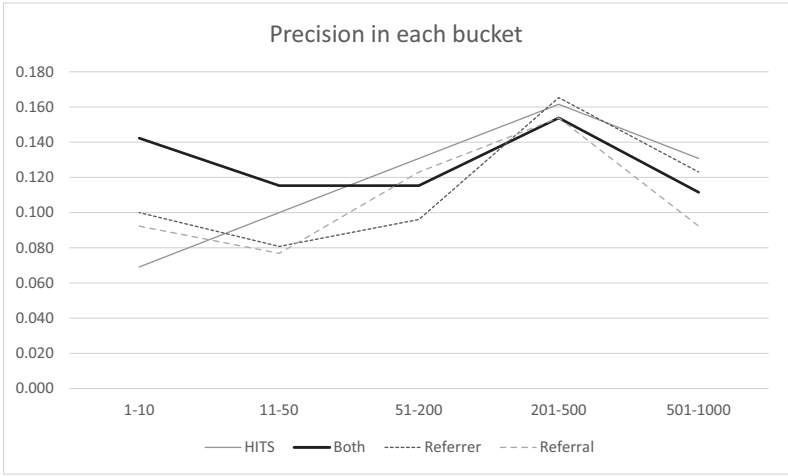

Figure 5.4: Results for all queries.

Bucket	HITS	Both	Referrer	Referral
1–10	0.175	0.225	0.188	0.175
1 – 10	0.069	0.142	0.100	0.092
11 – 50	0.100	0.115	0.081	0.077
51 – 200	0.131	0.115	0.096	0.123
201 – 500	0.162	0.154	0.165	0.154
501 – 1,000	0.131	0.112	0.123	0.092

Table 5.2: Precision of each bucket through all queries.

can be classified into easy queries and specialist queries: “Postal service privatization”, “France trip”, “The Sagrada Familia” and “Fish called by different names in life stage” are not specific queries, and “Parkinson’s disease”, “Game theory”, “Compiler” and “Machine learning” are specialist queries which mean academic or technical terms. Figure 5.5 and Figure 5.6 show the results for two types of query. In the easy query case, the ratio of correct pages is high: 0.35 through four queries. When the search task was easy, the search result contained many useful pages for novice users. Each method had similar precision in the buckets of the upper part of the ranking. The **both** method found more correct pages than other methods in the middle part of the ranking. The **referral** method had findings throughout the ranking. The **referrer** method performed worse than the original HITS. In the specialist query case, the total number of correct pages in the dataset was small, with

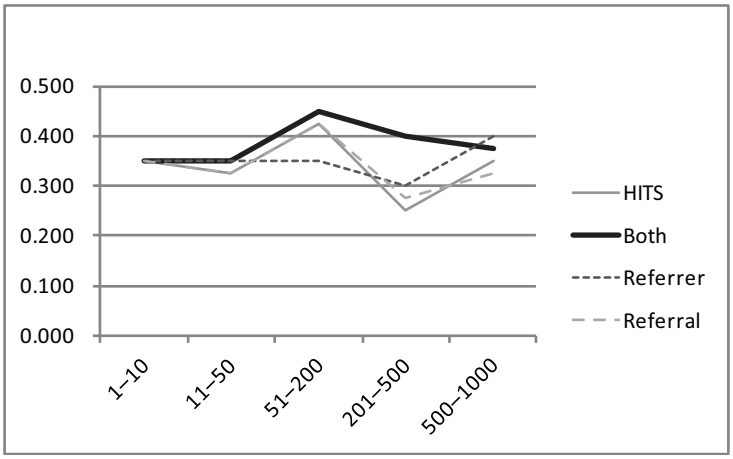

Figure 5.5: Results for non-academic queries.

a ratio of just 0.09. It is assumed that pages on specialist topics are commonly not written in a language easy enough for novice users to understand.

On the whole, the **both** method, which uses both referral and referrer diversity, works well, especially when the query is specialist. The HITS algorithm ranked correct pages in the middle part of the ranking, and the two **referral** or **referrer** methods used the original adjacency matrix $\boldsymbol{A}$. They were influenced strongly by original HITS result. Although the total number of correct pages was small, the **both** method ranked many correct pages in the top part of the ranking and not many of them in the bottom.

5.4 Discussion

Throughout the whole evaluation, the proposed method could find more suitable pages for novice users with a high degree of accuracy, but that precision depends on the query.

Looking at the case of eight original queries, the proposed method works better when the query is specific. For example, when the query is "machine learning", the proposed method finds three suitable documents, while basic HITS cannot find any suitable documents. The detail of pages judged as correct document are online dictionary sites, or introductory works written by academic society. Dictionary sites and encyclopedia sites are frequently linked by many individual personal blogs. The author of each blog has different interests. Authors that are incidentally interested in "machine learning" will create links to those pages. Another day, they write about other interests, and link other interesting pages. Then these pages have links not only to "machine learning" pages. As looking forward all queries,

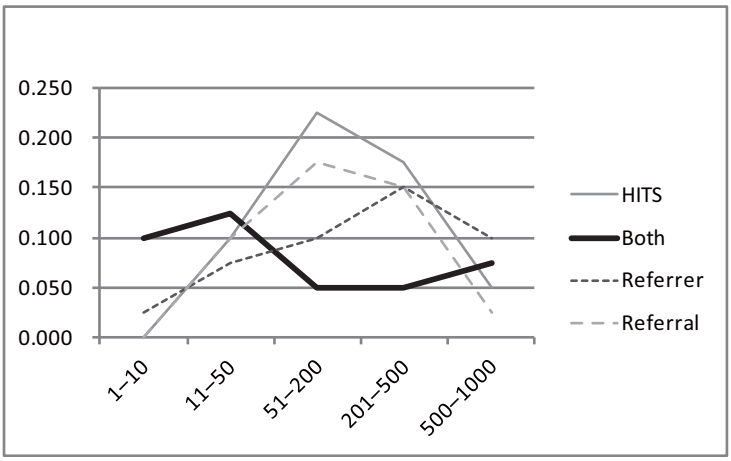

Figure 5.6: Results for specific queries.

our method tend to have high accuracy in the case of unique noun queries and concrete queries. The correct pages have two deferent types: documents containing definitions such as dictionary pages, introductory guide and expositions, and event documents such as diaries, blogs and so on. When the queries are unique nouns or concrete terms, the search result contains many documents containing definition. Our method worked well relatively in definitive documents. In the case of both types of pages are mixed, our method could not evaluate the relative usefulness of them. For instance, sometimes someone's diary article is more useful for novice users, but the popularity and the diversity of reference is smaller than difficult documents which contain definition. On the other hand, there are many popular blogs which have no useful information. It would appear that the method proposed should be used to compare documents of the same type.

On the other hand, pages judged as not suitable are documents deemed too difficult, low-quality pages, spam and pages not relevant to the query. In this case, some social bookmarking sites were found in the top part of the basic HITS ranking. They do not contain useful information. These sites are strongly connected with themselves by internal links. They are characterized by a high score for both Hubs and Authority. The basic HITS algorithm is weak to such kind of link structures. Pages from social bookmarking services feature a similar design template. Our method estimated their topics as similar to each other.

The proposed method did not work well when the query was easy. The number of correct pages found is same to HITS in the top part of the ranking. Our method found more correct

page in the middle and bottom part of ranking. In this task, the number of pages suitable for novice users is inherently large. When the query is general, relevant pages are general too. For instance, when the query was "France trip", each method yielded results mostly containing content from major travel agency sites, all of which are about hotels, touristic hot spots or itineraries. When a judgment is determined only by the relevance between the query and page, HITS-based algorithms may not be suitable even if it is expanded. There is a possibility that the diversity factor causes reverse effect, that is, the document linked only by documents about trips may be relevant to the trip. Then, non-diverse tight links provide higher relevancy.

The graph of basic HITS algorithm has its peak in the middle of the ranking, where it scores good pages. In the bottom of the ranking, there are many pages linked by few pages. Most of these pages are spam pages, low-quality pages, or minor pages. In the top part of the ranking, a lot of individual pages of major online service sites appeared. They are linking to each other. Some of them have no content inside, *e.g.*, private pages in social bookmarking websites, message pages in online fora with the aim to drive communication, product introduction pages of big company websites and so on. These pages are perceived as spam, or are not relevant to the query. Two variant methods which take odd of two diversity could not overcome it.

The effectiveness of the proposed method is not so high in the middle part and bottom part of the ranking. It can be said that there is a limitation of link analysis approach to improve finding suitable pages for novices. The participants found many correct pages that are linked by few pages. The pages that are useful but not popular still exist. Our improvement cannot help to find such kind of pages, that is, the link is the only key to the computation usefulness. Without a reasonable amount of links, link analysis methods have no power. It should be used complementarily with content-based approaches to cope with such situations.

Even though our method works better than original HITS algorithm in this experiment, its accuracy is not high enough yet. The method has to be optimized by fixing and tuning parameters. For instance, the methods used in this experiment are not tuned, so it did not take into account the weight of propagation score and diversity score, the distribution of diversity score, the tuning of LDA and so on.

5.5 Summary

We proposed a diversity-based Web ranking method that expands on the HITS algorithm to include two diversity-based hypothesis: 1) that a page linking diverse Authorities is a

valuable Hub, and 2) that a page linked by diverse Hubs is a valuable Authority. The objective of the diversity-based HITS algorithm is to enable not limited specialist users but general users to find suitable documents, that is, documents that are useful for novice users. We defined diversity as how the topic of each document in a set of documents is different from the topics of the other documents. We call the diversity of pages linking to the page referrer diversity and the diversity of pages linked by the page referral diversity. We expanded the HITS algorithm by replacing the adjacency matrix with two diversity-based matrixes. The proposed methods were compared with the original HITS algorithm by their authority scores in terms of finding useful pages for novice users. The method that uses both referral and referrer diversity could rank more good pages high, especially when the search query was specific.

As future work, we intend to expand the diversity-based methods further. Our method abandoned many factors to simplify the model. Of course, the method itself is built around the idea of diversity, not popularity, so it is necessary to focus on the number of linking documents, the amount of information, and the power of influence pages. Moreover, diversity can be defined not only from the topic of pages: for example, we can define it as authors' property, sentiment of documents, temporal-spatial metadata, and so on. We will tackle these issues with additional diversity-based methods.

CHAPTER 6

Conclusions

6.1 Summary

This thesis has discussed the Web information search and diversity analysis techniques on the social Web information. Nowadays, the World Wide Web has aspects of social media, *e.g.*, communication media, commercial media, and so on. Existing Web search systems rank Web pages by the term frequency in their contents and the number of hyperlinks from/to them. However, these features do not sufficiently reflect the social aspects of the current Web. In this thesis, we propose three search methods which consider social backgrounds of Web pages, *e.g.*, who focuses the page, why they focus the page and how diverse they are. These methods are based on the discussions about the three research topics below:

- **Search-by-Reaction:**
  Web Search Using People's Reaction Terms in Twitter
  Traditional Web search engines cannot retrieve the pages relevant to the emotional keywords such as "cute" and "interesting" sufficiently. This is because such emotional words are mainly used to describe reactions to Web pages, and the original pages do not necessarily contain them. To solve the problem, we propose a new concept for advanced Web search, Search-by-Reaction, and develop a search method under the concept. The method complement the contents of Web pages with the description of reactions to the pages. To extract the reaction description, our method focuses on the hyperlinks between the original pages and social networking services which contains many reaction description. Our evaluation indicated the benefit of our method.

The proposed method could find suitable pages more accurately than existing search method for the two types of queries (*i.e.*, the impression query and the topic query).

- **Diversity-Based Credibility Analysis:**
  Can Diversity Improve Credibility of User Review Data?
  Review is a type of reaction. After the retrieval of reviews, the next problem is their credibility. To solve the problem, we propose methods to estimate the credibility of a reviewer and a group of reviewers. The reviewers with higher credibility estimate more precisely the quality of items. Our methods are built on two assumptions about links between reviewers and items: 1) A reviewer who has reviewed many and diverse items has high credibility and 2) a reviewer group which consists of many and diverse reviewers has high credibility. To verify the assumptions, we conducted experiments with a movie-review data set. The results showed that the diversity of items and reviewers was an effective feature for estimation of the credibility of reviewers. One finding is that the type of the diversity is important to estimate group's credibility. The diversity measure based on *variety* and *balance* can find credible groups. In contradistinction to this, the diversity measure based on *disparity* does not work well to find credible groups.

- **Diversity-Based HITS:**
  Web Page Ranking by Referrer and Referral Diversity
  More generally, hyperlinks are a type of positive reaction from the source page to the target page. Some well-known page ranking algorithms, namely PageRank and HITS, are based on this observation, and pages receiving more in-links have better chance to gain better score. However, even if two pages receive the same number of in-links, there is a major difference between the pages linked from pages of few homogeneous topics and the pages linked from pages of many and diverse topics. To rank general Web pages considering this difference, we propose a method focusing on the diversity of links. We propose two types of link diversity, namely referral diversity and referrer diversity. Our method calculates scores based on the above two types of diversity and our extended HITS algorithm. The experimental results showed that our method is more useful than the original HITS algorithm to retrieve Web pages suitable for novice readers.

6.2 Future Directions

There are still several research topics that need to be further explored in future work. One research topic is to merge the three topics described above. The concepts of the users' credibility and the diversity among users can help improving the Search-by-Reaction model. The credibility of the user who takes a reaction is important to estimate if the page is actually impressive or not. The diversity of the users who take reaction may be a good guide to estimate the generality of the reaction on controversial pages. The rest of proposed methods also have points to be improved. The current goal of diversity-based credibility analysis method is to find a credible reviewer or group. It can be used in item search, such as to find SF movie only acclaimed by SF experts. Diversity-Based HITS must be improved on its method and application. Accuracy enhancement and more efficient calculation is necessary. Using content-based methods and machine learning techniques may enhance the search of suitable information for novice users. Focusing on another types of diversity (*e.g.*, temporal diversity, sentiment diversity, *etc.*) makes the method applicable to more varied situations.

Using the social Web information furthermore, not exclusively informatics, but also social science, cognitive science and so on may also be important. For instance, the concept of independency, which is the knowledge of the group dynamics area can be discussed as a substitute of the diversity. The analysis of social big data (*e.g.*, SNSs' log, CGM data, *etc.*) with an interdisciplinary approach should improve the information search on the future social Web.

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, pages 5–14. ACM, 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, pages 40–46. IEEE Computer Society, 2011.

[3] X. Amatriain, N. Lathia, J. M. Pujol, H. Kwak, and N. Oliver. The wisdom of the few: A collaborative filtering approach based on expert opinions from the web. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 532–539. ACM, 2009.

[4] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[5] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 43–50. ACM, 2006.

[6] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In *Proceedings of the 16th International Conference on World Wide Web*, pages 501–510. ACM, 2007.

[7] L. Barbosa and J. Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 36–44. Association for Computational Linguistics, 2010.

[8] A. Bifet and E. Frank. Sentiment knowledge discovery in twitter streaming data. In *Discovery Science*, pages 1–15. Springer, 2010.

[9] K. Bischoff, C. Firan, W. Nejdl, and R. Paiu. Can all tags be used for search? In *Proceeding of the 17th ACM conference on Information and knowledge management*, pages 193–202. ACM, 2008.

[10] G. Capannini, F. M. Nardini, R. Perego, and F. Silvestri. Efficient diversification of web search results. *Proceedings of VLDB Endow.*, 4(7):451–459, 2011.

[11] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 335–336. ACM, 1998.

[12] D. Carmel, N. Zwerdling, I. Guy, S. Ofek-Koifman, N. Har'el, I. Ronen, E. Uziel, S. Yogev, and S. Chernov. Personalized social search based on the user's social network. In *Proceeding of the 18th ACM conference on Information and knowledge management*, pages 1227–1236. ACM, 2009.

[13] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, pages 675–684. ACM, 2011.

[14] 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, pages 239–248. ACM, 2009.

[15] A. Dong, R. Zhang, P. Kolari, J. Bai, F. Diaz, Y. Chang, Z. Zheng, and H. Zha. Time is of the essence: improving recency ranking using twitter data. In *Proceedings of the 19th International Conference on World Wide Web*, pages 331–340. ACM, 2010.

[16] N. Fuhr. Probabilistic models in information retrieval. *The Computer Journal*, 35(3):243–255, 1992.

[17] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 2009.

[18] 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, pages 576–587. VLDB Endowment, 2004.

[19] T. Haveliwala. Topic-sensitive pagerank: a context-sensitive ranking algorithm for web search. *Knowledge and Data Engineering, IEEE Transactions on*, 15(4):784 – 796, 2003. july-aug.

[20] D. Helic, M. Strohmaier, C. Trattner, M. Muhr, and K. Lerman. Pragmatic evaluation of folksonomies. In *Proceedings of the 20th International Conference on World Wide Web*, pages 417–426. ACM, 2011.

[21] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? In *Proceedings of the international conference on Web search and web data mining*, pages 195–206. ACM, 2008.

[22] J. Hu, G. Wang, F. Lochovsky, J. Sun, and Z. Chen. Understanding user's query intent with wikipedia. In *Proceedings of the 18th International Conference on World Wide Web*, pages 471–480. ACM, 2009.

[23] Y. Jin, R. Li, K. Wen, X. Gu, and F. Xiao. Topic-based ranking in folksonomy via probabilistic model. In *Artificial Intelligence Review*, pages 1–13. Springer Netherlands, 2011.

[24] A. Josang, R. Ismail, and C. Boyd. A survey of trust and reputation systems for online service provision. In *Decision Support Systems*, volume 43, pages 618–644. Elsevier, 2007.

[25] M. Kato, H. Ohshima, S. Oyama, and K. Tanaka. Can social tagging improve web image search? In *Web Information Systems Engineering-WISE 2008*, pages 235–249. Springer, 2008.

[26] Y. Kawai, Y. Fujita, T. Kumamoto, J. Jianwei, and K. Tanaka. Using a sentiment map for visualizing credibility of news sites on the web. In *Proceedings of the 2Nd ACM Workshop on Information Credibility on the Web*, WICOW '08, pages 53–58. ACM, 2008.

[27] C. Khoo, N. Hamzah, S. Chan, and J. Na. Sentiment-based search in digital libraries. In *jcdl*, pages 143–144. ACM Press, 2005.

[28] S. Kim and E. Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, pages 1367–1373. Association for Computational Linguistics, 2004.

[29] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, Sept. 1999.

[30] T. Kumamoto and K. Tanaka. Proposal of impression mining from news articles. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 901–910. Springer, 2005.

[31] K. Kuroda and M. Hagiwara. An image retrieval system by impression words and specific object names-iris. *Neurocomputing*, 43(1-4):259–276, 2002.

[32] R. Lempel and S. Moran. The stochastic approach for link-structure analysis (salsa) and the tkc effect. *Computer Networks*, 33(1 – 6):387–401, 2000.

[33] T.-Y. Liu. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331, 2009.

[34] X. Liu, W. B. Croft, and M. Koll. Finding experts in community-based question-answering services. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, pages 315–316, 2005.

[35] Y. Lu, M. Castellanos, U. Dayal, and C. Zhai. Automatic construction of a context-aware sentiment lexicon: an optimization approach. In *Proceedings of the 20th International Conference on World Wide Web*, pages 347–356. ACM, 2011.

[36] Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi. Exploiting social context for review quality prediction. In *Proceedings of the 19th International Conference on World Wide Web*, pages 691–700. ACM, 2010.

[37] A. E. Magurran. Measuring biological diversity. *African Journal of Aquatic Science*, 29(2):285–286, 2004.

[38] C. Manning, P. Raghavan, H. Schütze, and E. Corporation. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, UK, 2008.

[39] J. J. McAuley and J. Leskovec. From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 897–908. International World Wide Web Conferences Steering Committee, 2013.

[40] D. W. McDonald and M. S. Ackerman. Just talk to me: A field study of expertise location. In *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work*, CSCW '98, pages 315–324. ACM, 1998.

[41] K. McNally, M. O'Mahony, B. Smyth, M. Coyle, and P. Briggs. Towards a reputation-based model of social web search. In *Proceeding of the 14th international conference on Intelligent user interfaces*, pages 179–188. ACM, 2010.

[42] E. Minack, G. Demartini, and W. Nejdl. Current approaches to search result diversification. In *Proceedings of The First International Workshop on Living Web at the 8th International Semantic Web Conference (ISWC)*, Oct. 2009.

[43] J. Na, C. Khoo, S. Chan, and N. Hamzah. Sentiment-based search in digital libraries. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 143–144. ACM, 2005.

[44] S. Nakamura and K. Tanaka. Video search by impression extracted from social annotation. In *Web Information Systems Engineering-WISE 2009*, pages 401–414. Springer, 2009.

[45] M. Nakatani, A. Jatowt, H. Ohshima, and K. Tanaka. Quality evaluation of search results by typicality and speciality of terms extracted from wikipedia. In *Proceedings of the 14th International Conference on Database Systems for Advanced Applications*, DASFAA '09, pages 570–584. Springer-Verlag, 2009.

[46] K. A. Nolan and J. E. Callahan. Beachcomber biology: The shannon-weiner species diversity index. In *Proceedings of Workshop ABLE*, volume 27, pages 334–338, 2006.

[47] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[48] S. E. Page. *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton University Press, 2008.

[49] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of LREC 2010*. European Language Resources Association (ELRA), 2010.

[50] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, Jan. 2008.

[51] 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, pages 275–281. ACM, 1998.

[52] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 109–126. Gaithersburg, MD: NIST, January 1995.

[53] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, pages 851–860. ACM, 2010.

[54] X. Sha, D. Quercia, P. Michiardi, and M. Dell'Amico. Spotting trends: The wisdom of the few. In *Proceedings of the Sixth ACM Conference on Recommender Systems*, RecSys '12, pages 51–58. ACM, 2012.

[55] B. Smyth, E. Balfe, J. Freyne, P. Briggs, M. Coyle, and O. Boydell. Exploiting query repetition and regularity in an adaptive community-based web search engine. In *User Modeling and User-Adapted Interaction*, volume 14, pages 383–423. Springer, 2004.

[56] A. Stirling. A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, 4(15):707–719, 2007.

[57] J. Surowiecki. *The wisdom of crowds*. Anchor, 2005.

[58] Y. Takahashi, H. Ohshima, M. Yamamoto, H. Iwasaki, S. Oyama, and K. Tanaka. Evaluating significance of historical entities based on tempo-spatial impacts analysis using wikipedia link structure. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, HT '11, pages 83–92. ACM, 2011.

[59] H. Tong. Fast random walk with restart and its applications. In *In ICDM ' 06: Proceedings of the 6th IEEE International Conference on Data Mining*, pages 613–622. IEEE Computer Society, 2006.

[60] E. Voorhees and D. Harman. Overview of the ninth text retrieval conference (trec-9). *Voorhees et Harman*, 2000.

[61] J. Wang and J. Zhu. Portfolio theory of information retrieval. In *Proceedings of the 32nd international ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 115–122. ACM, 2009.

[62] K. Wang, Z. Ming, X. Hu, and T. Chua. Segmentation of multi-sentence questions: towards effective question retrieval in cqa services. In *Proceeding of the 33rd international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 387–394. ACM, 2010.

[63] X. Wang, L. Zhang, F. Jing, and W. Ma. Annosearch: Image auto-annotation by search. In *Proceeding of IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 1483–1490. IEEE Computer Society, 2006.

[64] X. Wei and W. B. Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 178–185. ACM, 2006.

[65] Y. Xia, L. Wang, K. Wong, and M. Xu. Sentiment vector space model for lyric-based song sentiment classification. *Proceedings of of the Association for Computational Linguistics. Columbus, Ohio, USA: ACL-08*, pages 133–136, 2008.

[66] Y. Yanbe, A. Jatowt, S. Nakamura, and K. Tanaka. Can social bookmarking enhance search in the web? In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 107–116. ACM, 2007.

[67] Y. Yanbe, A. Jatowt, S. Nakamura, and K. Tanaka. Towards improving web search by utilizing social bookmarks. In *Web Engineering*, pages 343–357. Springer, 2007.

[68] D. Yimam-Seid and A. Kobsa. Expert-finding systems for organizations: Problem and domain analysis and the demoir approach. *Journal of Organizational Computing and Electronic Commerce*, 13(1):1–24, 2003.

[69] S. Yu, D. Cai, J.-R. Wen, and W.-Y. Ma. Improving pseudo-relevance feedback in web information retrieval using web page segmentation. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pages 11–18. ACM, 2003.

[70] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, pages 403–410. ACM, 2001.

[71] J. Zhang, Y. Kawai, T. Kumamoto, and K. Tanaka. A novel visualization method for distinction of web news sentiment. In G. Vossen, D. Long, and J. Yu, editors, *Web Information Systems Engineering - WISE 2009*, volume 5802 of *Lecture Notes in Computer Science*, pages 181–194. Springer Berlin Heidelberg, 2009.

[72] L. Zhuang, F. Jing, and X. Zhu. Movie review mining and summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 43–50. ACM, 2006.

Acknowledgements

I would like to express my sincere gratitude to my thesis supervisor Professor Katsumi Tanaka for his always valuable advice, constructive discussions and warm encouragement. He always supports my research through his great research vision and foresight.

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 Professor Kazutoshi Sumiya at University of Hyogo for his valuable discussions and suggestions.

Special thanks should be given to Emeritus Professor Yoshifumi Masunaga at Ochanomizu University and Professor Martin J. Dürst at Aoyama Gakuin University. They were supervisors of my master's course and undergraduate days. They ongoingly give helpful advisement for my research.

I wish to thank Professor Osami Kagawa at Osaka Gakuin University, Professor Keishi Tajima, Associate Professor Adam Jatowt, Specific Associate Professor Yoko Yamakata, Specific Associate Professor Hiroaki Ohshima, my seniors in Tanaka laboratory Dr. Yusuke Yamamoto, Assistant Professor Takehiro Yamamoto and Assistant Professor Makoto P. Kato for their good comments and suggestions during the laboratory meeting.

I want to thank the secretaries of Tanaka laboratory, Ms. Ikebe, Ms. Sato, Ms. Tabata and Ms. Shiraishi for their kindly help. I would like to thank my colleagues in Tanaka laboratory, especially Dr. Kosetsu Tsukuda, Mr. Tomohiro Manabe and Mr. Kazutoshi Umemoto for their fruitful discussions.

Finally, I am deeply grateful to my father Tokutami Shoji and Keihou Shoji for their unlimited love and strong support for my study. Thank you.

Publications

Journal Papers

1. 荘司慶行, 田中克己, “印象語をクエリとするリアクションに基づくウェブ情報検索”, 情報処理学会 論文誌ジャーナル *JIP* 「情報爆発」特集号, 52(12), pp. 3515 - 3526, 2011.

2. 増永良文, 石田博之, 伊藤一成, 伊藤守, 清水康司, 荘司慶行, 千葉正喜, 長田博泰, 福田亘孝, 正村俊之, 森田武史, 矢吹太朗, “WikiBOK を用いた社会情報学の知識体系構築実験”, 日本データベース学会論文誌, vol.12, No.1, pp. 133 - 138 , 2013.

3. 荘司慶行, 田中克己, “リアクションサーチ:リアクションをクエリとするウェブ情報検索”, 日本データベース学会論文誌, Vol.10, No.1, pp.43 - 48, 2011.

4. 増永良文, 石田博之, 伊藤一成, 伊藤守, 清水康司, 荘司慶行, 高橋徹, 千葉正喜, 長田博泰, 福田亘孝, 正村俊之, 矢吹太朗, “集合知アプローチに基づく知の創成支援システム WikiBOK の研究・開発”, 日本データベース学会論文誌, Vol.10, No.1, pp.7 - 12, 2011.

International Conference Papers

1. Yoshiyuki Shoji, Makoto P. Kato, Katsumi Tanaka, “Can Diversity Improve Credibility of User Review Data?”, *Proc. of the 6th International Conference on Social Informatics (SocInfo 2014)*, pp. 244 - 258, Barcelona, Spain, November, 2014.

2. Yoshiyuki Shoji, Katsumi Tanaka, “Diversity-Based HITS: Web Page Ranking by Referrer and Referral Diversity”, *Proc. of the 5th International Conference on Social Informatics (SocInfo 2013)*, pp. 377 - 390, Kyoto, Japan, November, 2013.

3. 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 Proc. of the 10th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-10 Workshop)*, pp.243-249, Tokyo, Japan, June, 2013.

4. 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 Proc. of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-9 Workshop)*, pp. 202 - 207, Tokyo, Japan, June, 2012.

5. Yoshifumi Masunaga, Masaki Chiba, Nobutaka Fukuda, Hiroyuki Ishida, Kazunari Ito, Mamoru Ito, Toshiyuki Masamura, Hiroyasu Nagata, Yasushi Shimizu, Yoshiyuki Shoji, Toru Takahashi, Taro Yabuki, "Developing WikiBOK: A Wiki-based BOK Formulation-aid System", *Proceedings of 2011 First IRAST International Conference on Data Engineering and Internet Technology (DEIT 2011)*, pp. 960 - 963, Bali, Indonesia, March, 2011.

6. Yoshifumi Masunaga, Yoshiyuki Shoji, Kazunari Ito, "A Wiki-based Collective intelligence Approach to Formulate a Body of Knowledge (BOK) for a New Discipline", *Proceedings of the 6th International Symposium on Wikis (WikiSym 2010)*, pp. 1 - 9, Gdansk, Poland, July, 2010.

7. Yoshifumi Masunaga, Yoshiyuki Shoji, Kazunari Ito, "Collective Intelligence Approach for Formulating a BOK of Social Informatics, an Interdisciplinary Field of Study", *Proceedings of the 5th International Symposium on Wikis (WikiSym 2009)*,pp. 1 - 2,Orland, Florida, October, 2009.

Domestic Symposium and Workshops

1. Yoshiyuki Shoji, Katsumi Tanaka, "Diversity-Based Credibility Analysis of Social Review Data", *The 6th International Workshop with Mentors on Databases, Web and Information Management for Young Researchers (iDB 2014 Workshop)*, July, 2014.

2. Yoshiyuki Shoji, Katsumi Tanaka, "Page Ranking by Referrer-Diversity and Referral-Diversity", *The 5th International Workshop with Mentors on Databases, Web and Information Management for Young Researchers (iDB 2013 Workshop)*, July, 2013.

3. Yoshiyuki Shoji, Katsumi Tanaka, "Diversity-Based PageRank: A Ranking Method Based on the Diversity of In-link Pages", *The 4th International Workshop with Mentors on Databases, Web and Information Management for Young Researchers (iDB 2012 Workshop)*, July, 2012.

4. 荘司慶行, 田中克己, "リンク参照元ページ群の話題多様性に基づく被参照ページの品質推定", 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)* 会議録, A1-1, 2013.

5. 荘司慶行, 田中克己, "リアクションサーチ: リアクションをクエリとするウェブ情報検索", 第 *3* 回データ工学と情報マネジメントに関するフォーラム *(DEIM2011)* 会議録, F1-5, 2011.

6. 増永良文, 石田博之, 伊藤一成, 伊藤守, 清水康司, 荘司慶行, 高橋徹, 千葉正喜, 長田博泰, 福田亘孝, 正村俊之, 森田武史, 矢吹太朗, "WikiBOK を用いた社会情報学の知識体系構築実験", 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)* 会議録, C3-5, 2013.

7. 増永良文, 石田博之, 伊藤一成, 伊藤守, 清水康司, 荘司慶行, 高橋徹, 千葉正喜, 長田博泰, 福田亘孝, 正村俊之, 森田武史, 矢吹太朗, "知の創成支援システム WikiBOK における構造化オブジェクトの編集競合解決法", 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)* 会議録, C8-1, 2012.

8. 増永良文, 石田博之, 伊藤一成, 伊藤守, 清水康司, 荘司慶行, 高橋徹, 千葉正喜, 長田博泰, 福田亘孝, 正村俊之, 矢吹太朗, "集合知アプローチに基づく知の創成支援システム WikiBOK の研究・開発", 第 *3* 回データ工学と情報マネジメントに関するフォーラム *(DEIM2011)* 会議録, A2-5, 2011.

9. 荘司慶行, 伊藤一成, 増永良文, "集合知的手法による学際的学問分野の知識体系構築システムのプロトタイピング", 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)* 会議録, C1-5, 2010.

10. 増永良文, 荘司慶行, 伊藤一成, "集合知形成手法を用いた学際的学問分野の知識体系記述の試み", 第 *1* 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)* 会議録, C3-5, 2009.

11. 荘司慶行, 松田基弘, 伊藤一成, Martin J. Durst, "統合型メタ翻訳の提案", 電子情報通信学会データベース学会合同ワークショップ *(DEWS2008)* 会議録, I2-24, 2008.