

**Master Thesis**

**Forming Wisdom of Crowds  
by Visualizing Web Pages**

Supervisor    Professor Toru Ishida

Department of Social Informatics  
Graduate School of Informatics  
Kyoto University

**Naoya YOSHIMURA**

February 8, 2007

# Forming Wisdom of Crowds by Visualizing Web Pages

Naoya YOSHIMURA

## Abstract

The way a user interacts with information on the Web has changed along with the permeation of the Web into the general public. During the Web creation period, web pages were registered to portal sites in the form of links, and users visited these portal sites to look for information. This method functioned well when the amount of total information on the Web was relatively small, but as the amount of information grew, finding information by following the portal links alone would not guarantee the discovery of needed information. The search technique is widely used as a method of filtering user-requested information from a vast amount mixed information.

Services using what is called the “Wisdom of Crowds” have recently been gaining attention as a new trend on the Web. Wikipedia, an online encyclopedia which anyone can edit, is a prime example of high-usage information source created using the Wisdom of Crowds. Wikipedia contains expert-level knowledge covering a broad range of areas and categories, and it is free. Creation of such knowledge content is possible by enabling users to easily participate in the formation of Wisdom of Crowds.

However, the following two problems exist when such Wisdom of Crowds is formed:

1. Understanding the whole construct of the Wisdom of Crowds is difficult
2. Knowing the content area which needs contribution is difficult

This research aims to solve the above problems by assisting the formation of Wisdom of Crowds using PlainView, a Web page visualization system. PlainView can map a set of web pages onto a two-dimensional plane to generate a holistic view of those web pages. Existing concept of “complementary search” is applied to the system to visualize and discover the content area that requires user contribution. This research attempts to bring the target set of Wisdom of Crowds close to the superset of the Wisdom of Crowds by applying the complementary search technique. The approach is also evaluated.

For the first problem, understanding the grasp of the whole image of Wisdom of Crowds is assisted by using PlainView. The user collects Web pages from Web site where set wisdom such as Wikipedia has been used and Web page arbitrary set, and compares the views. And, it pays attention to each feature (close of the view and sparse part), and the whole image is understood.

For the second problem, understanding and discovering insufficient knowledge for the Wisdom of Crowds by comparing the views generated by setting a similar topic and the axis between two or more set. For instance, it thinks about the case to use the match-up set: Web and the object set: Wikipedia. At this time, it is clear that Web that is the match-up set has more various information than Wikipedia that is the object set. The user compare the views of both, and to be able to discover insufficient knowledge for Wikipedia.

Moreover, to distinguish close and a sparse part on the view when Web page set is done in the mapping the point of plotted should scatter to solve the second problem. It is necessary to set the word used for the axis well for that. It has been understood that devices of the setting of two or more words that become not only a single concretely word but also synonyms, and the consideration of literary words and the spoken language, etc. are necessary.

Contributions of this research are following two points.

1. Proposes the mechanism that the whole image of Wisdom of Crowds can be understood.
2. Support the discovery of contents for which the contribution is needed.

In the above-mentioned activity, the concept of existing Complementary Search was applied to a new area of Wisdom of Crowds formation. Concretely, the proposal technique was applied to Wikipedia as a real problem and the evaluation and discussion.

As a result, it has been understood that information linked with the activity by the real world (for instance, information such as the society activity, thesis information, the references, and manuals etc.) are lack in Wikipedia. Moreover, those lack information showed the discovery by comparing the views between two or more set such as Web and Wikipedia.

## Web ページの可視化を用いた集合知の形成

吉村 尚也

### 内容梗概

Web の一般化に伴って利用者と情報の関わり方も変化している。Web の創成期においては情報を発信した場合、ポータルとなるリンク集にそのリンクが登録されることによって利用者に認知されていた。この方法は Web 上の情報の総量が比較的少ない場合には上手く機能していたが、次第に総量が増加するに従ってリンク集で纏め上げることが困難になった。また、情報の量だけでなく利用者が求める内容も多種多様に拡散していった。そのような膨大で多種多様な情報の中から利用者の目的とする情報をフィルタリングする方法として今日では検索技術が広く利用されている。

そのような中、Web の新たな潮流として集合知を利用したサービスが注目されている。近年では個人の利用者が社会問題や TV 番組、各種商品といった日常で触れるさまざまなトピックに関して、ブログや掲示板、SNS などを通して自由に意見を述べる機会が増えている。その結果、Web 上には公式な情報だけでなく、利用者自らによって作成された様々な情報が日々蓄積されている。中でも、誰でもが編集可能であるオンライン百科事典の Wikipedia は、専門家水準の知識が蓄積されている事、多種多様な項目を有している事、項目の作成や利用には基本的に金銭的成本を必要としない事などから、利用価値の高い集合知の代表例と言える。これら Wikipedia の持つ特徴は、集合知の形成に多様な利用者が参加する事によって実現されている。

しかしながら、そうした集合知の形成時には次の 2 つの問題がある。

1. 集合知の全体像を把握することが困難
2. 貢献が必要とされているコンテンツの発見が困難

本研究では上記の問題を解決することに取り組む。具体的には、Web ページ可視化システムである PlainView を用いて、集合知の形成を支援する。Plain View は Web ページ集合を 2 次元平面上にマッピングして可視化することにより、多様な情報を含んだ集合の全体像を提供することを実現している。この PlainView に、既存の Complementary search の概念を適用することで、貢献可能なコンテンツの発見を試みる。本研究では Complementary search の定義における対象集合の集合知の全体像を、母集合の集合知の全体像に近づけることを目的として評価を行っている。

1.の問題に対しては PlainView を用いる事によって集合知の全体像の把握を補助している。利用者は Wikipedia 等の集合知が活用されている Web サイトと任意の Web ページ集合（本研究ではこれを母集合とし、適用例では Web 全体としている）から Web ページの収集を行い、互いのビューを比較することによって密や疎な部分といったそれぞれの特徴に着目し、全体像を把握できる。

2.の問題に対しては複数の集合間において、同様のトピック・軸の設定によって生成されたビューを比較することによって、目的の集合に足りない知識を発見することが可能となる。例えば、母集合に Web、対象集合に Wikipedia を用いた場合を考える。このとき、母集合である Web のほうが対象集合である Wikipedia よりも多様な情報を有していることは明らかである。利用者は両者のビューを比較し、Wikipedia に足りない知識を発見することが可能となる。

以上のようにして、2つの問題の解決に取り組んだ。しかしながらビューの生成時においては一度の操作で利用者にとって有益なビューが得られるとは限らない。そこで、利用者はシステムと対話的にビューの編集を行う事が可能となっている。さらに、ある利用者によって作成されたビューは他の利用者と共有する事も出来る。その結果、利用者は互いのビューを利用することによってビュー作成時の負担を軽減することができる。

また、2.の問題を解決するためには Web ページ集合をマッピングする際に、ビュー上で密と疎な部分が判別できるようプロットされた点が散らばる必要がある。そのためには、軸に用いる単語を上手く設定しなければならない。具体的には単一の単語だけでなく同義語となる複数の単語を設定したり、書き言葉や話し言葉を考慮したりするなどの工夫が必要であることがわかった。

本研究の貢献をまとめると以下の2点となる。

1. 集合知の全体像を把握できる仕組みを提案
2. 貢献が必要とされているコンテンツの発見を補助

上記の活動の中において、既存の Complementary Search の概念を集合知形成という新たな領域に適用した。具体的には、実問題として Wikipedia に提案手法を適用し評価と考察を行った。

その結果、Wikipedia では現実世界での活動とリンクする情報（例えば学会活動、論文情報、リファレンスやマニュアルといった情報）などが不足していることが分かった。それらの不足情報は Web と Wikipedia など複数の集合間におけるビューの比較によって発見できることを示した。

# Forming Wisdom of Crowds by Visualizing Web Pages

## Contents

<b>Chapter 1 Introduction</b>	<b>1</b>
<b>Chapter 2 Relation between Wisdom of Crowds and User</b>	<b>5</b>
2.1 Feature of Wisdom of Crowds .....	5
2.2 Problem when Wisdom of Crowds is formed .....	6
<b>Chapter 3 Visualizing Web Pages</b>	<b>9</b>
3.1 Related Works to Visualizing Web Pages .....	13
3.2 User-Centered Approach to Visualizing Web Pages.....	16
3.3 PlainView.....	20
<b>Chapter 4 Complementary Search</b>	<b>23</b>
4.1 Explanation of Complementary Search .....	23
4.2 Comparison with Existing Retrieval .....	24
4.3 Application Image to Wikipedia.....	29
<b>Chapter 5 Application Example to Wikipedia</b>	<b>34</b>
5.1 Application Example 1: “OS” .....	34
5.2 Application Example 2: “Google” .....	35
5.3 Application Example 2: “Agent $\cap$ AI” .....	36
<b>Chapter 6 Discussion</b>	<b>38</b>
6.1 Support for User when Wisdom of Crowds is Formed .....	38
6.1.1 Grasp of Whole Image of Wisdom of Crowds .....	39
6.1.2 Discovery of Contents to be able to Contribute .....	39
6.2 Efficient Discovery Technique .....	41
6.2.1 Words used for Axis .....	41
6.2.2 Interaction with System .....	42
6.3 Use of Wisdom of Crowds .....	43
<b>Chapter 7 Conclusion</b>	<b>45</b>
<b>Acknowledgments</b>	<b>47</b>



# Chapter 1 Introduction

The relation between user and Web information has changed along with the generalization of Web. At the creation period of Web, sent information was limited to contents that a laboratory of the university and some special users made. Therefore, the gross weight of information was relatively little compared with present, and the Web page was acknowledged to the user by the URL's being registered in links that became portal. When the user received intelligence, it received intelligence by settling, selecting the appropriate one according to each purpose from among raised links, and tracing the link. This arranged, systematized information on each Web site, and an individual Web page was recognition of part on the Web site rather than contents. Settling by links and raising appeared by developing and generalizing Web various Web pages, and increased gradually also the gross weight became difficult though this method functioned well when the gross weight of information on Web was comparatively little.

At this time, a general user etc. who do not have the enterprise, government and municipal offices, and expertise in addition to the above-mentioned caller are enumerated as for sending information on the Web site. When the concern for Web rose, and the social influence came to be recognized, the caller of various information came to use Web positively. As a result, the number of Web sites not only increased but also it became the one with various contents of an individual Web site. It can be said that this came to maintain contents that one Web site is various while only the content along the purpose with a specific there is Web site till then was often maintained. As a result, the mechanism that the Web site was able to be inspected crossing came to be needed when thinking that it wanted information on a certain topic.

Additionally, not only sent information but also the content that the user requests has diffused variously. The search technique is used widely and generally as a method of filtering information assumed to be user's purpose from among such huge, various information today. It can be said that this is a technology that collects all information by the robot type retrieval , for example,

Google, and attempts the match with information that the user requests.

Service using Wisdom of Crowds is paid to attention as a new current of in such and Web. Recently, the chance to express one's views freely through Blog, the BBS, and SNS, etc. has increased for various topics that an individual user touches in daily life (news and various products). Especially, it can be said that Wikipedia that everyone can edit is an example of the representative of Wisdom of Crowds with high utility value. It can be said that such present Web has the following possibilities in potential.

### **1. Grasp of whole image concerning topic**

The person in the world can search for the entire tendency with what idea might there for the topic that is by efficiently collecting information on Web. Subjective individual opinions that exist on Web are actually efficiently collected, and the research used for the risk management of marketing and the enterprises of the information gathering and the marketing research, etc. at the merchandise purchase has been done before[1][2][3][4]. The whole image concerning the topic can be easily understood spending neither money nor time by it was necessary to do a large-scale questionnaire survey using information on Web so far.

### **2. Offer of various views with a large amount of sample**

The offer of various views is from among independently a large amount of sample to can the extraction of information that the user wants through media. Sending information is limited to the newspaper and media of radio before Web spreads, and information has been sent by the people very limited. Even if information that those media pass on is the re-composition of the reality of the continuousness of the choice, and doesn't have a special intention, I cannot help entering it by the production person's value judgment.

There are major information, minor information, small number of informational and large number of informational in the composed knowledge base depending on various information sending means (Blog, BBS, and SNS on the other hand). It is thought that the user can mention the view of various the one by touching various information on Web.



Figure 1.1: Services using Wisdom of Crowds<sup>1</sup>

As mentioned above, it is assumed to be a purpose of this research to support the formation of Wisdom of Crowds by using the technique of making the whole image of Web page group visible for gaining power the Web service that uses Wisdom of Crowds. It proposed the method of supporting the tool rice field attention to Wikipedia as the example of the representative that used Wisdom of Crowds, and making the article in Wikipedia, and the evaluation and consideration were done.

It was not as more difficult as present Web at an initial stage than present that the number of articles understood the whole image few in Wikipedia. However, the number of articles that Wikipedia possesses is exceeding 300,000 now, and the grasp of the whole image is difficult in man's ability. It is useful that the user who wants to contribute to Wisdom of Crowds named Wikipedia by describing the article obtains the whole image of the article though the purpose can be achieved by using the retrieval to discover the article.

For the user who wants to contribute to Wisdom of Crowds, it is a purpose to describe the article on the area based on knowledge and the experience that I possess. Such knowledge and the experience are organically related as a set of various information. The user comes to be able to offer information that consists of own knowledge and experience for Wisdom of Crowds according to

<sup>1</sup> <http://ja.wikipedia.org/>, <http://ja.wikipedia.org/>, <http://okwave.jp/>

the difference if the set of the knowledge that such I possess and information of experience can compare Wisdom of Crowds with information that has. However, even if the user can discover information by the pinpoint, it is difficult only in existing information retrieval to understand the whole image. The user can be said that the page that should be inspected is huge and he or she is difficult when thinking from the current state that keeps increasing now though it is not impossible the presumption of the whole image of information as the image in the head by inspecting the page concerning the topic one by one.

It grappled with this problem by using the system named PlainView[5] described in the above-mentioned in this research. The concept named Complementary Search that is the information retrieval to discover information that doesn't exist described in Chapter 4 on that is used. Afterwards, the example of application to Wikipedia that is actually the example of the representative of Wisdom of Crowds in Chapter 5 is shown. And, the discussion of the support of the formation of Wisdom of Crowds based on the application result is described in Chapter 6. Finally, the conclusion of this research is described in Chapter 7.

# **Chapter 2 Relation between Wisdom of Crowds and User**

Service that uses Wisdom of Crowds, for example, Wikipedia by which the user participates in making contents on Web as described in Chapter 1 is paid to attention. The user can share knowledge and the experience by using Wisdom of Crowds. This chapter considers why Wisdom of Crowds coming is paid attention besides by thinking about the feature of Wisdom of Crowds from various angles though doesn't know. In addition, not only the advantage but also the disadvantage exists in Wisdom of Crowds. It pays attention to the part of the disadvantage, and the problem with which it should grapple by this research is clarified.

## **2.1 Feature of Wisdom of Crowds**

It was the main that only some users sent information before, and many other users inspected information. On the other hand, not only some users but also various users participate in sending information, and my knowledge and experience are being offered to other users in Wisdom of Crowds. Various users' having contents that are more various than offered information in Wisdom of Crowds according to participation in sending information the limited user named the enterprise and the group becomes possible.

Moreover, Wisdom of Crowds possesses an advantage from respect of the cost. A big cost is needed by being in case of various one the contents when the enterprise and the group try to make contents. However, Wisdom of Crowds doesn't need the cost from based on the motive that it wants to offer the user's knowledge and experience to the generation of contents. It is easy to understand as the idea of this by the example of Wikipedia. A part of publisher made, and the encyclopedia had been sold before. Because the encyclopedia became an amount of tens of volumes in the one that contained a lot of contents, the publisher had to mobilize a lot of number of men for the making. Therefore, a lot of costs were needed by making the encyclopedia. And, the user

who used the encyclopedia had to pay the publisher a large sum of value.

On the other hand, Wikipedia can be used free of charge if there is an environment that can be connected with Web. Moreover, special contents not published for the convenience of space and the cost (Or, publish) and local contents are published in the encyclopedia before. Moreover, a big feature is always a description named Web of latest information from the environment with high character at once.

Of course, contents with low quality exist in the inside because various users are participating in making contents, too. However, even if only the same contents are compared by the time series because it is possible to edit it while being exposed to a lot of eyes of the user, and supplementing the description mutually, it is possible to keep being refined at any time.

The mechanism that can be understood intuitively for various users not organized at this time to participate in contents is important though making contents that cover a wide area (long tail) without needing an economical cost becomes possible by using Wisdom of Crowds. For instance, it can be said that the spread rolls the ordinary user by offering the mechanism that can be intuitively understood and it accelerated though it was the one that can be achieved even if the service that Blog and SNS are providing is existing Web.

Wikipedia that the composition such as consumption and making contents by the user is clear is paid to attention in this research though some services using Wisdom of Crowds exists. The item that is necessary is retrieved as well as a general Web site or it only has inadvertently to trace the directory and to obtain it for the consumption of contents.

## **2.2 Problem when Wisdom of Crowds is formed**

When various advantages make contents while it is, it has some problems in Wisdom of Crowds as described in a passage. In it, it is one of the features of Wisdom of Crowds, and the thing that various users are participating in the generation of contents is a cause. The situation that information is not systematized at the same time is developed though various users participate in Wisdom of Crowds and Wisdom of Crowds's obtaining diversity, too became

possible. This is because the method of classifying information and the method of Categoraz are different depending on the user.

It is a mechanism that one item can belong to two or more categories in Wikipedia corresponding to this now. As a result, the advantage that it is possible to discover it by two or more methods when the user as the consumer of information wants to discover information has been brought. However, the situation of not understanding how the written item is related when contents are described at the same time each other easily begins also to see production as for it.

The problem of taking it up chiefly has the following two.

1. It is difficult to understand the whole image of Wisdom of Crowds.
2. Do it only have to be made contents to belong to which category?

The problem of one first of all is described. When contents that the user wants to make are limited, it is good, and causes the demand that it wants to understand contents concerning it covering it when it tries to make a certain contents. For instance, when the item of artificial intelligence is described, the user can often describe the item like the agent, the multi agent, and the cognitive science, etc. that relate to it. If the item group that centers on the item of artificial intelligence at such time, and relates to it can be presented covering it, it becomes easy to make contents or more. The user can acquire the whole image of the item that relates on Wikipedia from his knowledge and the viewpoint of experience.

Doing the description that sees easily becomes possible by uniting the appearances of words and phrases and sentences after the item is made if the relating item can be confirmed to it beforehand when the same item of departure is made. In addition, lack and repetition can be excluded when describing it and expect it.

Next, the problem of two is described. The fragmentary information is raked up as a result of obtaining in present Wikipedia by the retrieval, and it is not possible to do this year when the user imagines the whole image from them by own ability. It is thought that it is difficult to set up knowledge and the experience from a micro like this aspect in the system and to build it in Wisdom

of Crowds. Focus is done from macro aspect to a gradually micro aspect by the interaction of the user and the system by the method of offering the whole image in this research, it works in the category to which the user is appropriate, and it works on making the item. It is thought that it becomes easy to judge to which category to only have to belong from the interrelation if the relating item group can be presented covering it.

## Chapter 3 Visualizing Web Pages

The note in this chapter's explaining making of information used to achieve the purpose of this research visible, and making it to visible is described. The early research on visible making information in WWW is enumerated on that. And, it explains the concept of making of the man center that is the architecture of PlainView used by this research visible.

First of all, making to visible is to make it see in some shape as for information that cannot be originally looked directly. Technology (scientific visualization) that expresses it in shape to understand the simulation result or the range and the distribution of the attribute ground easily has been mainly researched in the field of the science and technology calculation so far. As for the characteristic of such making to visible, same information has been pointed out that the point that our easiness degree of acknowledgment is different according to the selection of the presentation and the fact sheet reality from early time. Therefore, it came to be valued to save findings abundant as the technique of making to visible of best information was able to be selected according to kind of a variety of information and task. For instance, the load of the expansion and the retrieval work of the resource that the user memorizes or can process is reduced, the information pattern is supported detection specific, and the method of making to visible has been devised by the purpose of offer of an easy-to-use interface by an operation the information sight directly. There is a basic data structure like the temporal[6], the layered structure[7], and the link structure[8][9], etc. as a method of making to visible. And, there is a system to take a general view, and to search for a lot of data interactively [10][11][12][13][14].

Recent years to support the user who looked for a target Web page from among a large amount of Web page that existed on Web . Various systems that described in a passage were developed. Because the feature of making of information visible targeted set of the abstract realities, and made information without a numeric attribute and a geographic attribute visible effectively, application to a wide field became possible. For instance, making information

visible is used as a document retrieval in an electronic library, a support function of writing, and a user interface for the program development. Moreover, expressed information is required to change from various sides and the abstraction levels dynamically by not a static snap shot making to visible but the interaction.

Especially, when the user necessarily has neither clear knowledge nor the definition of the retrieval object like information retrieval beforehand, it is important to improve the retrieval demand gradually by feedback from the system.

As mentioned above, the visible information making is a direct program, and personally a technology (the document file, the directory, and the Web space) that presents information that cannot be seen in eyes in a comprehensible form. The characteristic of making to visible is a point that our easiness degree of acknowledgment is different according to the selection of the presentation and the fact sheet reality, and can be used to understand user's information when the best mode of expression is taken even by same information. Especially, the user can capture the feature, offers an ideal feedback information without reading details of the document by using visible making information, and even a target document is interactively good at the field of information retrieval at the navigation from among a large amount of document. As for information that exists on Web, after the Internet spreads, it increases day by day, and efficiently obtaining target information becomes impossible from among that. The key role is borne on the problem, and the visible information making can be called a technology to be worthy.

Technically, the following technique, algorithms, and the applications are the main themes.

- Modeling of target information for making to visible
- Design of sight fact sheet reality used for making to visible
- Mapping from modeled information to sight data
- In the operation information retrieval to sight data

The visible information making is very effective to obtain target information from the barrage of information. It doesn't put a strain in the user by gripping

the feature without taking a general view of the whole, and reading details and information can be retrieved. In the following, it explains what approach taken about the technical theme described in the above-mentioned in information retrieval.

First of all, there are the following seven in a sight fact sheet reality.

- **Linear(1-D Linear)**

The arrangement of the element of one dimension like the text and the source code, etc. hits this.

- **Two dimensions(2-D Map)**

It is a group of the element with the area like the map. When the whole image is presented as a map, it is used.

- **Three dimensions(3-D World)**

It is a gathering like the realities of the real world of that ..possession.. as for the extension of three dimensions.

- **Temporal**

The one that especially concerns time series in linear element. A time change in information is expressible by the combination with two dimensions and three dimensions.

- **Multi-dimensional**

Set of elements like class of relational database with attribute of n piece.

- **Tree**

Set where each element has parents element only of one excluding element that becomes root. The structure of the structure and the Web site of the organization and each categories are peeled off ..making of the filesystem of sets of documents put together and computers.. visible.

- **Network**

Set with the arbitrary connection relationship between elements that are more general than tree. It is suitable for making of the one like the reference relation of the document, the link structure of the Web site, and the history etc. of the transition between user's Web pages visible.

Next, the case of information retrieval, especially two of the following is paid to attention, and the mapping to the sight data is done.

- **Presence of the relations between two or more information**

The relations between two or more information (the reference relations between documents, the link structures between Web pages, and the transition histories between Web pages) are extracted, information is done, and the mapping is done to the network where the relations between the node and information are assumed to be an edge. When an individual relation such as the community's discovering and making of traffic on the network visible is valued, it is used.

- **Content of information**

The sight data is made from the content being written in the document. There is a technique of clustering of the content base in the typical one. In clustering, the similar level between documents is calculated by making the feature vector in which the document is first characterized from the content of each document, and calculating the distance of the feature vector. And, when making it to visible, the document with large similar level is arranged mutually to be near. The technology that pays attention to the content being written like this is used when searching for the information space or navigating. The target one is extracted additionally in the document, and there is a method of making to visible based on the extracted data, too. Statistic of transition of the approval rating of the achievement of prices of merchandise, sales, and the company and Cabinet is extracted from the newspaper and the Web document, etc. , and for instance, it is remarkable in recent years by doing the mapping domestically and started up the workshop in the graph.

And, there are the following seven in the operation to sight data.

- **Overview**

Understands whole image of general view

- **Zoom**

Expands element and part where attention should be done

- **Filter**

Removes an element not interesting

- **Details-on-demand**

Gets in-depth information from individual element and specific group

- **Relate**

Display the relations between elements

- **History**

Manages user's operation history

- **Extract**

Record session management and the retrieval result

It is advocated as a pattern in which making to visible succeeds in information retrieval, "It takes a general view first of all, and on-demand and make it in detail after the zoom and the filter", and has the function by the information retrieval system that uses making to visible a lot of facts[15]. The technology such as a general view, a zoom, filters, and detailed on-demands is an important operations for retrieving the barrage of information. For instance, information (focus) to which it pays attention arranges at the center of the screen, zooms the understanding of details, is misinterpreted the surrounding of information at the center like the fish-eye lens, and the focus+context technique for offering context (context) of extent in which the page to which it pays attention is not obstructed is adopted by a lot of researchers. [11][8][16][17] The focus+context technique is known to be effective to refer to information interesting as global information for the browsing and the navigation is maintained by a necessary details degree.

It introduces the system developed for visible making information in WWW at the following.

### **3.1 Related Works to Visualizing Web Pages**

WEBSOM[12] is a system that makes information in the Internet visible by the Self Organization map. Figure 3.1. It is an example of making about 30,000 articles on Newsgroup that is by peel WEBSOM visible. The feature of the document is expressed by the key word set, and a similar document is arranged in neighborhood by the learning algorithm. The word that characterizes the

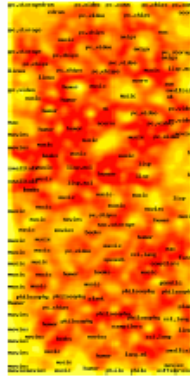


Figure 3.1: WEBSOM[12]

cluster that is the group of a similar document is selected, displayed automatically, and the density of the document is made to correspond to the color light and shade.

AdunaAutoFocus<sup>2</sup> clusters and offers the result of retrieving the file of the Web page and the local. Figure 3.2 is a result of the use of the user of the retrieval word such as “personal”, “knowledge”, and “management” and the retrieval. The document related to each retrieval word is put together in the cluster and displayed.

Grokker<sup>3</sup> retrieves the Web page and local files such as Yahoo and ACM Digital Library and Amazon Books based on the key word, and makes the retrieval result a map in shape like Figure 3.3. Each topic is brought together, and in the retrieval result, you on the view shows and each topic and the square show the Web page. And, two or more you that show the subtopic further are

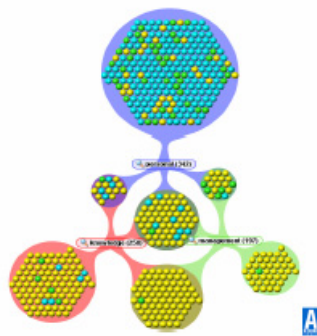


Figure 3.2: Aduna AutoFocus<sup>2</sup>

<sup>2</sup> <http://aduna.biz/products/autofocus/personal/index.html>

<sup>3</sup> <http://www.grokker.com>



Figure 3.3: Grokker<sup>3</sup>

included in the circle, and the whole can be seen in doing the expansion reduction and details be seen.

KartOO<sup>4</sup> is a meta search engine retrieved to all almost major search engines such as AltaVista and Teoma except Google. The feature of KartOO is to display the retrieval result in the map like Figure 3.4 where Flash was used. A lot of pages corresponding to the retrieval result are displayed in the map, and the result in which the evaluation is high can be distinguished according to the size of the page. The key word of each page of the retrieval result is displayed left. It is possible to search by narrowing it to the page in which it is interested in the selection of the key word. The one typical in the key word is displayed on the map, and the page related to the key word is arranged near the key word.



Figure 3.4: KartOO<sup>4</sup>

<sup>4</sup> <http://www.kartoo.com>

### **3.2 User-Centered Approach to Visualizing Web Pages**

Making the man center visible is to make Web page set concerning a certain topic visible by user's original standard, and to present the whole image from which the user requests the topic. It is difficult to satisfy the user's peculiar demand in an existing technology that only gives the topic by the user and makes it to visible. Then, the man center to describe such a problem as follows is approached.

First of all, the standard that is called an axis by the user is made to be specified to present the whole image that the user's peculiar demand is answered. The Web page lines up by the axis that the user specified, and can offer the whole image that the user requests as a result. The system doesn't completely think the Web page of the axis that the user freely specifies to be able to line up easily by the automatic operation. Then, the user will freely edit the result of the return of the system.

If the result of the line of the system is wrong, the user corrects it, and an important page for the user does the edit work that the emphasis display is made to be done. The system offers various functions for the user to edit it easily, and the user completes the whole image that he requests by doing the system and the interaction.

Concretely, the following four points are sequentially executed.

- Modeling of target information for making to visible
- Mapping from modeled information to sight data
- Sight fact sheet present used for making to visible
- Operation to sight data

It explains the technique to achieve the presentation of the whole image on the Web page that uses the above-mentioned axis at the following.

First of all, the object of making to visible is Web page concerning the topic that the user specifies. The user freely specifies the axis, and the Web page lines up according to the axis. Because the user can freely specify the standard, various standards are expected to be specified. Therefore, statistic and the feature word included in the Web page do not limit to the element with a specific it is, and use the majority of the word included in the Web page. All

words of the noun included in the Web page, the compound noun, the verb, and the adjective are concretely targeted in making to visible. In addition, the Web page is thought that the numeric representation and the name of a place (the charge, length, and weight) are also effective to line up. Then, these numeric representations are targeted in making to visible.

Next, the mapping to the sight data is to arrange it in the axis by which the user specifies the Web page. It is necessary to concern centering on the page to arrange it centering on the Web page and to quantify it. It is necessary to request strength of the relations between the word and the Web page that shows the axis in a word. Then, the following method is used as a method of the mapping.

- **Mapping that uses appearance frequency of word**

The word and the Web page are thought that the relation is high by the inclusion of a lot of words that the word that the user specifies is on the Web page. For instance, the Web page concerning the topic "Iraq war" is assumed and when word "Agreement" is included more than word "Opposite" when lining up with the axis "Agreement  $\longleftrightarrow$  Opposite", it is the Web page that "Opposite" and the relation are high of "Agreement" and the relation if "Opposite" is more high, and oppositely than "Agreement". And, the mapping to the sight data is done by arranging the Web page in one with a high more relation.

- **Mapping that uses numeric representation**

It thinks about the mapping that uses the numeric representation (the amount of money and the approval rating included in the Web page). Wanting the inclusion of the date expression in the numeric representation, the line with the axis by the sending time of the Web page, and person's Web page are used and the Web page concerning a certain event is used to line up with the axis concerning the date of generation of the event to line up with the axis concerning the person's date of birth.

- **Mapping that uses name of a place**

The method of doing the mapping to the sight data is devised by using the name of a place included in the Web page. Because the name of a place is a

character string, it converts it into the numerical value by using the dictionary etc. prepared beforehand.

Two dimension plane is composed of the axis that the user specified. The point scatters by the Web page that the mapping is done on two different axes composing two dimension plane of those axes, and the whole image that is called a view in shape like the map is offered. The point on the view shows the Web page. The axis that the user specified composes a spindle and a horizontal axis.

However, it is not easy to think the system generates such a view with the automatic operation. Then, the system and the user repeat the interaction in this research and the view is made. The system takes the standpoint only of supporting it. The system generates the view with the axis that the user specified first. The user completes the view by obtaining feedback making the view a starting point, and repeating the system and the interaction. As a result, it is thought that a more accurate view can be made, and the original view that the user requests in addition can be made.

The system configuration is described about visible making the Web page system (PlainView) that uses it by this research at the following.

Plain View can take a general view of the whole image on the page that exists on Web at one view by expressing the Web page as one point in two dimensions when the topic and the axis are input. The feature of Plain View is three of the following.

- 1. The view is made with user's original axis.**
- 2. A view different according to the switch of the axis is offered.**
- 3. The view is made by the interaction between system and user.**

The view that Plain View makes generates a view different depending on not the uniform one but the user. Therefore, it is thought that it is difficult for the system to make the view by the automatic operation completely. Then, the view is made while talking with the user with the system in Plain View. The flow to user's obtaining the view is as follows.

#### **i. Collection of Web pages**

When the topic that the user wants to examine is input, Web page

collection request is transmitted to the system. The system that receives the request collects Web pages concerning the specified topic from WWW, and stores it in the data base.

## **ii. Preprocessing of Web page mapping**

Collected Web pages are assumed to be a preprocessing to do the mapping on the axis, and the extraction of the numeric representation like the deletion of the HTML tag, the morphological analysis, the date, weight, and length, etc. and the extractions of the feature word are done. Extracted result of the morphological analysis and numeric representation and feature word are stored in the data base.

## **iii. Mapping on axis on Web page**

When the user specifies two axes after the preprocessing ends, the mapping request on the axis is transmitted to the system. The mapping does each Web pages collected on the specified axis to the system that receives the request.

## **iv. Presentation of view**

Two dimension plane is composed of two axes that the user specified that the mapping to the feature space ends, and it is presented to the user as a view.

## **v. Interactive view edit between user and system**

The user freely does the edit work, and can make the requested view for the view presented by the system. To preserve the made view, the preservation request is transmitted to the system. The system that receives the request stores information on the view in the data base.

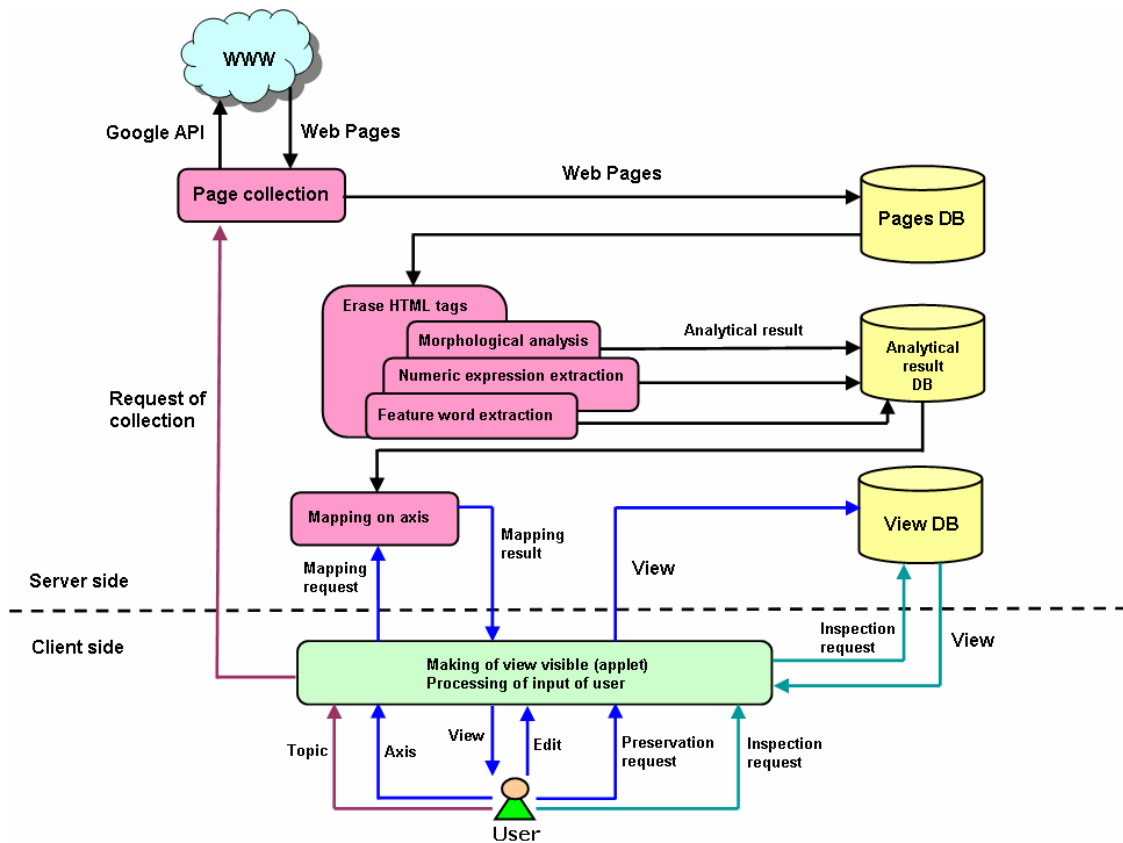


Figure 3.5: System configuration of PlainView[5]

### 3.3 PlainView

There is PlainView as a system that can collect information that fills the above-mentioned requirement covering it. PlainView is a system that plots the Web page as a point like two dimension plane that is called a view, and offers the user the whole image of Web page set. To obtain the view, the user gives the system the topic that he wants to be examining first of all. The system collects pages from Web based on the topic. If it wants to obtain the whole image from the sample more than at this time, the collected numbers of pages are reduced to enlarge the collected numbers of pages, and to complete the collection in short time or more. When the collection is completed, the user can generate the view. The collection of Web pages can begin the generation of the view immediately without collecting pages need not do at each use of the system, and to examine topics collected beforehand.

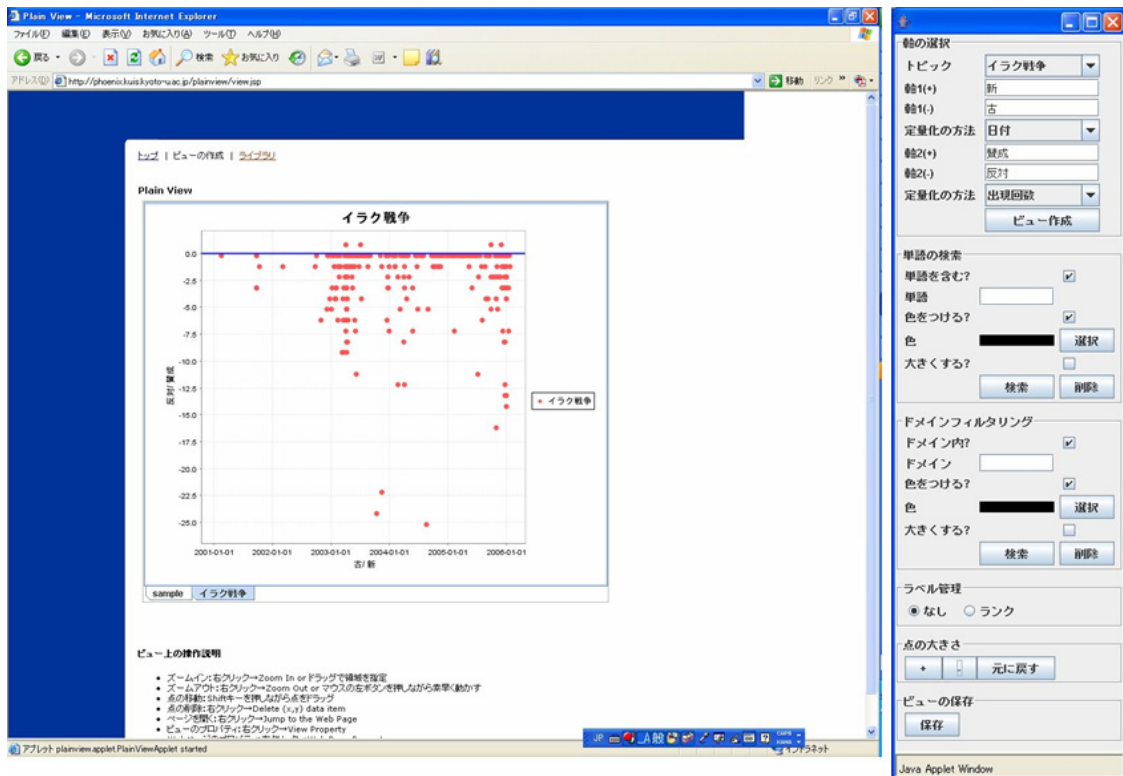


Figure 3.6: Example of using PlainView[5]

The view is generated by setting the evaluation method of use to a spindle and a horizontal axis chiefly. It is an arbitrary word that chiefly uses as an evaluation method. If he or she wants to examine the public opinion, the agreement and wording opposite are set, and if it attaches famously and he or she wants to examine it, the user uses goodness and the word badness of the commodity. As for the word, using two or more pieces becomes possible, and the generation of a flexible view according to each user's purpose is additionally possible with can the use of an arbitrary word.

In addition, the view that becomes a purpose because of an operation once when the view is generated is not necessarily obtained. Because view is generated with high flexibility, if word setting is not good, pages plot places are concentrate, inapposite place. The system can refine the view by operating it interactively to the user, and to obtain the view at which it aims as a result, is designed for that case. Concretely, even after having generated the view once, the user can replace and increase and decrease the word used for the axis for

the same view. As a result, to approach the aimed view, the user operates it while seeing the change in the view.

Figure 3.6 is a view about Iraq war. The time axis is taken in a horizontal axis, and agreeing or opposite word is used for the spindle. Whether at which there are a lot of Web pages where the opinion opposed the Iraq war is expressed, and time such a lot of opinions are expressed from this view can be read. It is achieved to offer the whole image that contains various information by making Web page set concerning the topic visible, and can understand the structure of the barrage of information intuitively in Plain View as shown in Figure 3.6.

## Chapter 4 Complementary Search

The problem described in Chapter 2 exists when contributing to Wisdom of Crowds (for example Wikipedia). Then, in this research when contributing to Wisdom of Crowds the method of presenting the user supplementary information corresponding to the purpose is used. It is assumed that the information gathering to support the formation of Wisdom of Crowds is called Complementary search in this research.

### 4.1 Explanation of Complementary Search

Complementary search is defined as follows.

#### Definition of Complementary search

Compare the match-up set of information concerning the topic with the object set, and present the user a sparse part of the object set for the match-up set.

Figure 4.1 shows the image of Complementary search. Information that should be presented to the user in Figure 4.1 is an area shown in light blue. In general, when the match-up set and the object set are compared, the match-up set elects a set that is bigger than the object set. For instance, match-up set is Web and object set is Wikipedia. At this time, it is thought that Web has the barrage of information more variously than Wikipedia. Therefore, an insufficient element can be discovered in Wikipedia though both are compared and it exists in Web. In figure 4.1, a light blue part is insufficient elements in Wikipedia though there is in Web.

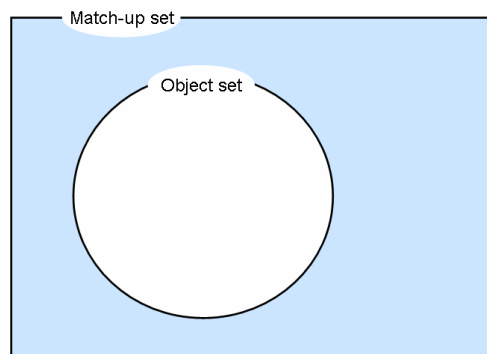


Figure 4.1: Match-up set and object set

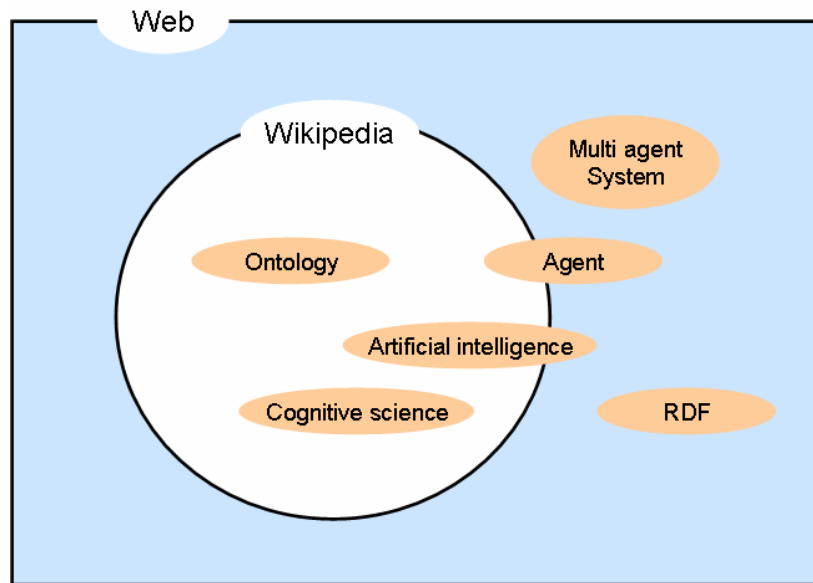


Figure 4.2: Example of applying Complementary Search

However, information that the user should understand of the collection is huge to discover contents to be able to contribute from among the area shown in light blue. Therefore, when the comparison result is presented to the user, it makes it to visible to make the whole image of collected information easy to understand. The user comes to be able to catch insufficient information for the object set efficiently by inspecting the comparison result of being made to visible (view).

It thinks about the case where the object set is assumed to be Wikipedia as an example as shown in Figure 4.2 by assuming the match-up set to be Web (The use of the technical dictionary is thought according to the application domain). At this time, when the information gathering of the topic (key word)"Artificial intelligence" is done, every information related to it is collected on Web. On the other hand, it is understood that the item is the Multi agent system "as soon as" Ontology "insufficient on Wikipedia and the user can understand as a result of the comparison.

## 4.2 Comparison with Existing Retrieval

"Existing information" is requested when past information retrieval and the

Table 4.1: Difference between information retrieval and Complementary search

	Information retrieval	Complementary search
<b>Purpose</b>	Discover “existing document” most related to the query	Discover "not existing (lacking) document" in an object set
<b>Input</b>	Query	Topic, Axis (Evaluation index for visualizing)
<b>Output</b>	List of related pages	View (Two-dimensional plane which did a plot of a pages according to the axis)
<b>Feature</b>	<ul style="list-style-type: none"> <li>•Revise the query by a provided result</li> </ul>	<ul style="list-style-type: none"> <li>•Generate the various views that accepted a demand of a user about one topic</li> <li>•Revise the view by a provided result</li> <li>•Discover the part which is sparse of a view by comparing a views</li> </ul>

difference of Complementary search are frankly described or "Information that doesn't exist" is requested. In the former, it doesn't limit to Wisdom of Crowds, it is used in various fields, and it is a purpose to obtain information wanting it. On the other hand, time when the user wants to contribute to Wisdom of Crowds chiefly is thought as a youth case as for the latter. To need when trying to contribute to Wisdom of Crowds is to say the item that has not been described in the domain where I have knowledge yet variously a variety of saying ,that is, "Information that doesn't exist". The method by the match of the character string cannot be used like existing information retrieval to discover information that doesn't exist.

Then, the discovery of information that depends on comparing not the content of information but whole images of information and doesn't exist is supported in Complementary search. For instance, it is thought that discovering information that doesn't exist in the object set becomes possible though each whole image is compared, and it exists from the difference point in the match-up set if it is possible to present it by making the whole image visible

respectively in the match-up set and the object set.

When existing information retrieval and the difference of Complementary search are brought together, it becomes as shown in Table 4.1. Information retrieval supports information that relates to Ceri and it is supported to discover information by matching the character string as stated above. It is supported to it to collect information by using the topic and the axis, and to discover information that doesn't exist from the whole image of the division obtained there in Complementary search. In the latter, various views corresponding to user's purpose can be obtained by changing the axis to the same topic.

The image of the relation between information and the user in past information retrieval and Complementary search is shown in Figure 4.3. In the former, information with high related level is acquired by filtering information. The editor of the article did this filtering, and past media were offering the user information that the importance degree had judged to be high widely. On the other hand, filtering by information retrieval has a big feature in the point of having improved flexibility from can the active decision of the user of the evaluation method more than media. However, there is a problem of biasing in several high ranks by matching the character string in many cases discovered information. That is, a page not so is to be likely hardly to be discovered though the page evaluated easily in the algorithm of information retrieval is discovered by a lot of users. The algorithm of information retrieval controls even the existence value on the page as a result.

In the latter, the point that focus can be done to an interesting part of the user after information on a certain topic is collected covering it, and they are made visible is a feature. Discovering information based on an own viewpoint after the whole image of information is understood becomes possible without controlling the algorithm of information retrieval from the user by doing so. Consequently, discovering information that doesn't exist in insufficient information for the object set, that is, the object set becomes possible by understanding the whole image though it exists in the match-up set.

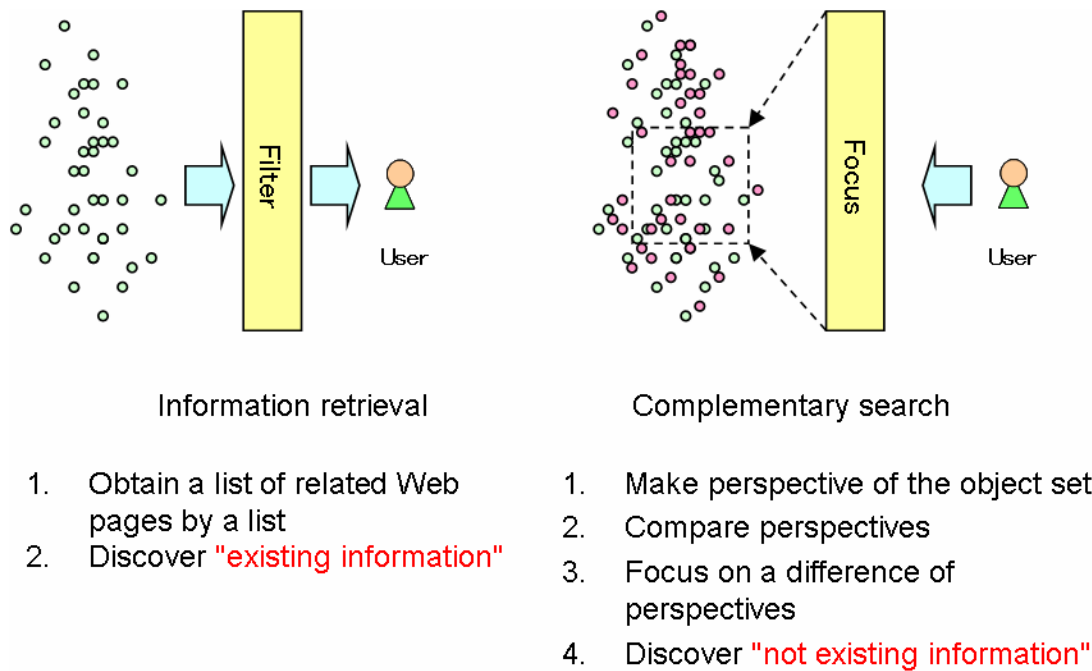


Figure 4.3: Image of relation between information and user

The retrieval performance is evaluated by using the relevance ratio and the recall ratio in an existing retrieval. The relevance ratio is a ratio of the document that suits the retrieval in the retrieval result. The recall ratio is a ratio of the document that suits the retrieval that was able to be retrieved. The standard of F-measure ( $=2 \times \text{recall ratio} \times \text{relevance ratio} \div (\text{recall ratio} + \text{relevance ratio})$ ) is used well so that the relevance ratio and the recall ratio are in the relation of the trade-off. Table 4.2 is a comparison of the evaluation indices.

On the other hand, because it cannot be clearly distinguished whether it is information because it is a purpose to discover information (contents for which the contribution is needed) that doesn't exist that the user is requesting in each item with Complementary search, it is not possible to evaluate it according to the relevance ratio and the recall ratio. However, it can be said that the evaluation of the view that the difference is caused in the distribution of the match-up set and the object set is high because it is a purpose to bring the knowledge set of the object set close to the knowledge set of match-up set in the point of distribution of information (It is useful for the user). If the difference is

Table 4.2: Comparison of method of evaluation

	Web page	Information retrieval	Wisdom of Crowds
method of evaluation	Page view, Page rank	Recall, Precision, F-measure	doesn't exist
Purpose	Discovery of <b>existing Web page</b> with high related level		Discovery of <b>knowledge that doesn't exist</b> to be able to contribute

caused, the user pays attention to the part, and catching information insufficient in the object set becomes possible.

Moreover, it can be said that it is a knowledge set that is better than the knowledge set of the object set (In the point of having approached the knowledge set of match-up set) when the view that the match-up set resembles the distribution of the object set is obtained for the word used for various axes.

In Figure 4.4. "coverage" and "unique" are defined to evaluates Wisdom of Crowds.

The ratio of the knowledge included in the object set of the match-up set is assumed to be coverage. The more it approaches 1.0 by the value of coverage, the more the object set (To the match-up set) has enough knowledge.

$$coverage = \frac{C}{M}$$

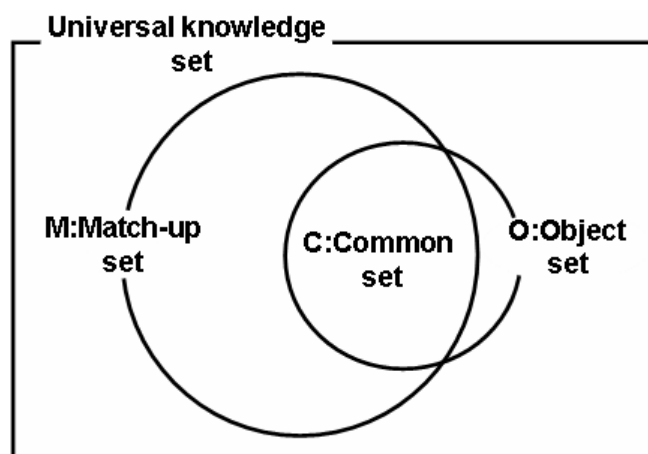


Figure4.4: Relation between match-up set and object set

The ratio of unique knowledge of the object set is assumed to be unique. The more it approaches 1.0 by the value of unique, the more the object set (To the match-up set) has unique knowledge.

$$unique = \frac{O - C}{O}$$

It is a target to bring coverage close to 1.0 in this research.

### **4.3 Application Image to Wikipedia**

The image of the application of Complementary search that makes to visible by PlainView to Wikipedia is shown in Figure 4.5. Here, pages are collected for the topic of artificial intelligence. Wikipedia has been selected as a collection target the object set of Web in the match-up set. It is expected that the difference is generated in the distribution of each view when making it to visible on that. For instance, the part in a green point that not is can be discovered though in the image of figure, there is a pink point in the upper part of the starting point. It becomes information of no existence in Wikipedia though this is in Web. What page an actual point is can change to an actual page by operating it on the point.

The user can discover the part where the distribution of the object set is sparse like this for the distribution of the match-up set. The user comes to be able to refine Wisdom of Crowds by describing and editing it if there is information to be able to contribute to Wisdom of Crowds by knowledge and the experience that the user has such a part.

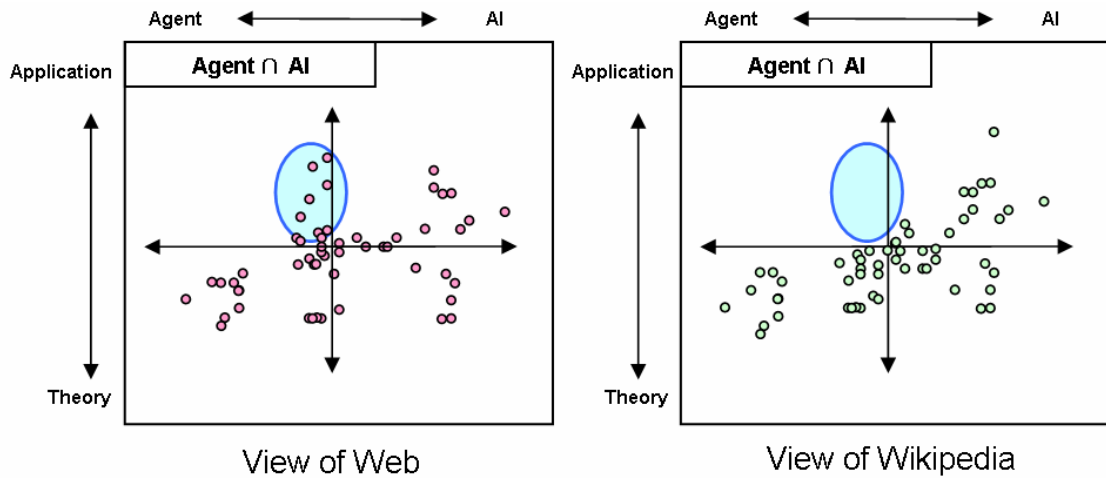


Figure 4.5: View of Complementary search

The flow chart of the procedure when the user generates the view on the system is shown in Figure 4.6.-4.8. Each work is described as follows.

**1. Selection of topic**

The user selects the topic of the field in which it can contribute to Wisdom of Crowds based on knowledge and the experience that oneself possesses.

**2. Retrieval of existing view**

It is retrieved whether other users have already made the view for the topic that the user selected. If something that has already been made exists, the user becomes possible by using the view the great omission of the following procedures, and can use the view efficiently.

**3. Specification of axis**

To generate the view, the user specifies the word used for the axis. Being able to use for the axis is a date, a prefecture name, and an arbitrary word. The one extracted on each Web page beforehand is used about the date and the prefecture name among these. The data base that did the morphological analysis to the Web page is used about an arbitrary word. The user can adjust the number of words if necessary because he or she can use two or more pieces about an arbitrary word.

**4. Generation of view**

The system plots the Web page like two dimension plane based on the

evaluation method specified for the axis by the thing that the user directs the generation of the view.

#### **5. The significant point is inspected**

As for the view generated by the system, the one assumed to be user's purpose by an operation once is not necessarily obtained. Therefore, it is confirmed whether (point plotted at a big position of the absolute value in a horizontal axis and the spindle) is inspected, and the content of the Web page and the plotted position are correct.

#### **6. The view is corrected**

When the position plotted as a result of the confirmation is amusing, the correction of the view is tried again by changing the word used for the axis.

#### **7. Comparison of views**

If all the significant points were able to be inspected, the user understands insufficient information for set wisdom by inspecting the view. The user compares two or more views. It is discovered what item you are relatively in a sparse part.

#### **8. Retouch and correction of item**

The item concerning the content of the discovered item is retouched and corrected. Insufficient information's for set wisdom coming to be added as a result, and enhancing the value become possible.

#### **9. Preservation of view**

In addition, when it is possible to contribute about other views, it retouches and it corrects it according to user's purpose. The made view is preserved in the data base if user's purpose is finally achieved and it ends. The use's of other users of the view becoming possible by preserving the made view in the data base, and reducing the load that rests upon the view making become possible.

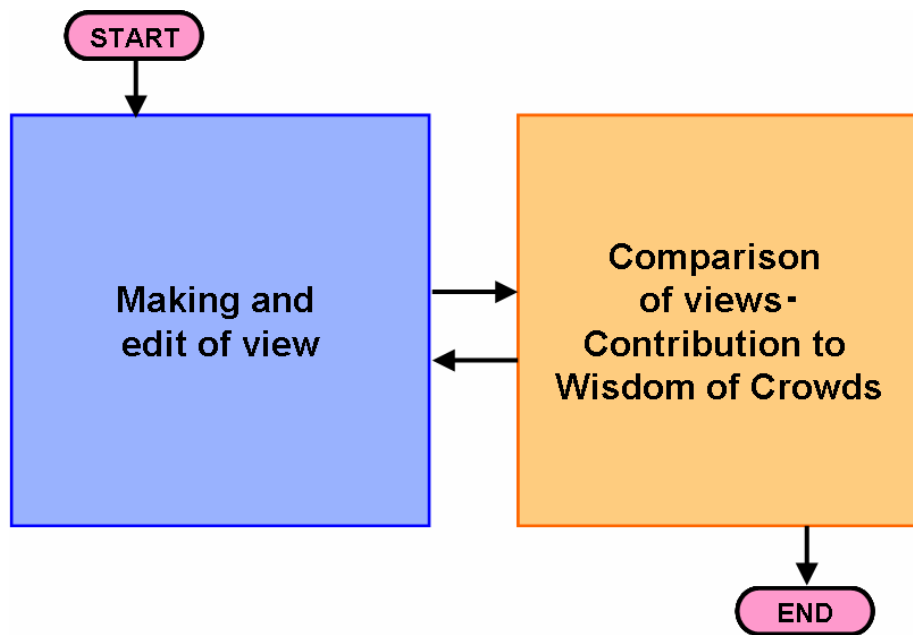


Figure 4.6: Whole image of user operation of Complementary Search

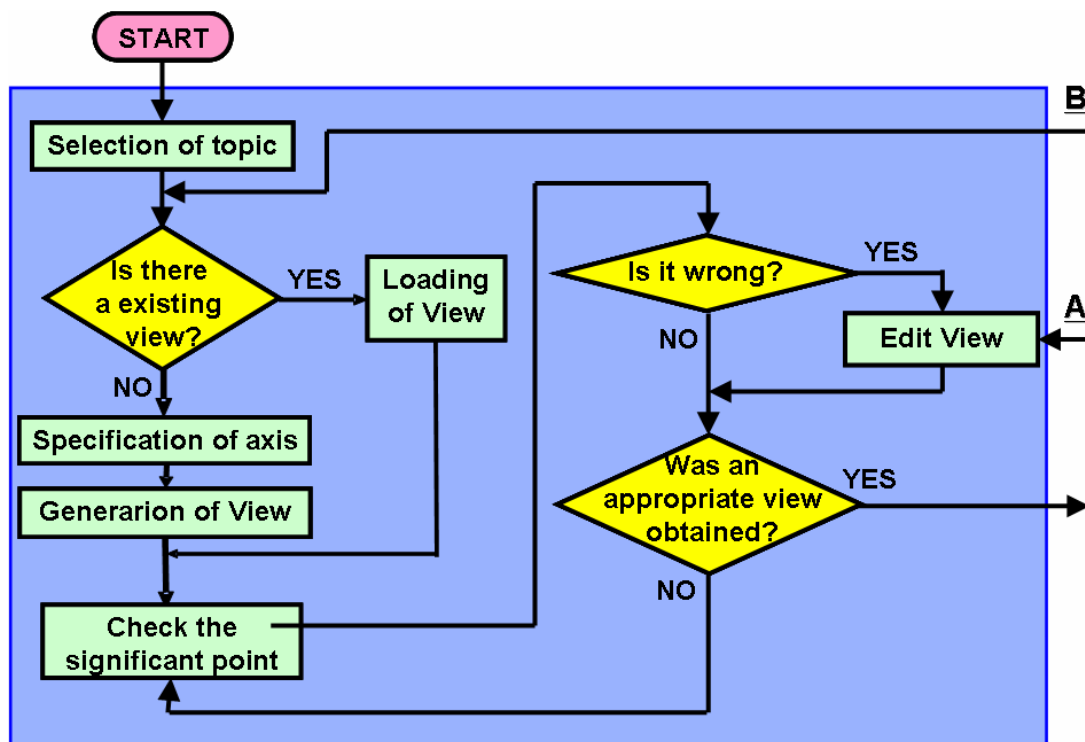


Figure4.7: Flow chart of making and edit of view

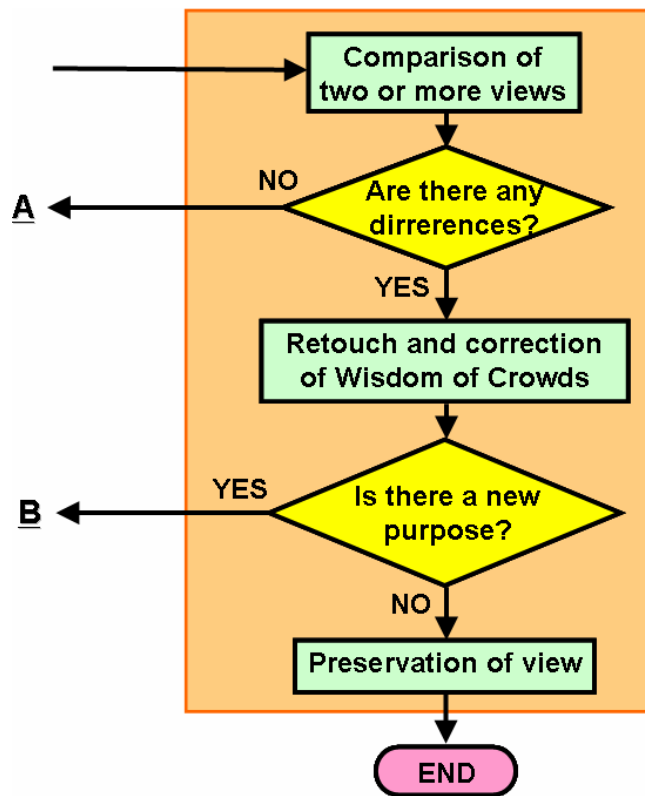


Figure4.8: Flow chart of Comparison of views and Contribution to Wisdom of Crowds

## Chapter 5 Application Example to Wikipedia

I show the example which applied the technique that I suggested below to Wikipedia. For an example, I have it of 3 that used "the OS", "Google", "Agent  $\cap$  AI" for a topic. In addition, as a result of having applied it, I show what kind of knowledge is not worthy of Wikipedia.

### 5.1 Application Example 1: “OS”

In Figure 5.1, the topic is a view of “OS”. A left view is a match-up set, and the collection object is Web. A right view is an object set, and the collection object is Wikipedia. The word used for the axis uses Windows or XP or 2000 or Vista or Me for a positive direction of a horizontal axis. Linux or UNIX or Minix or Redhat or Vine or GNU or BSD is used for a negative direction of a horizontal axis. Reference or command or installation is used for a positive direction of the spindle. Research or development or software is used for a negative direction of the spindle.

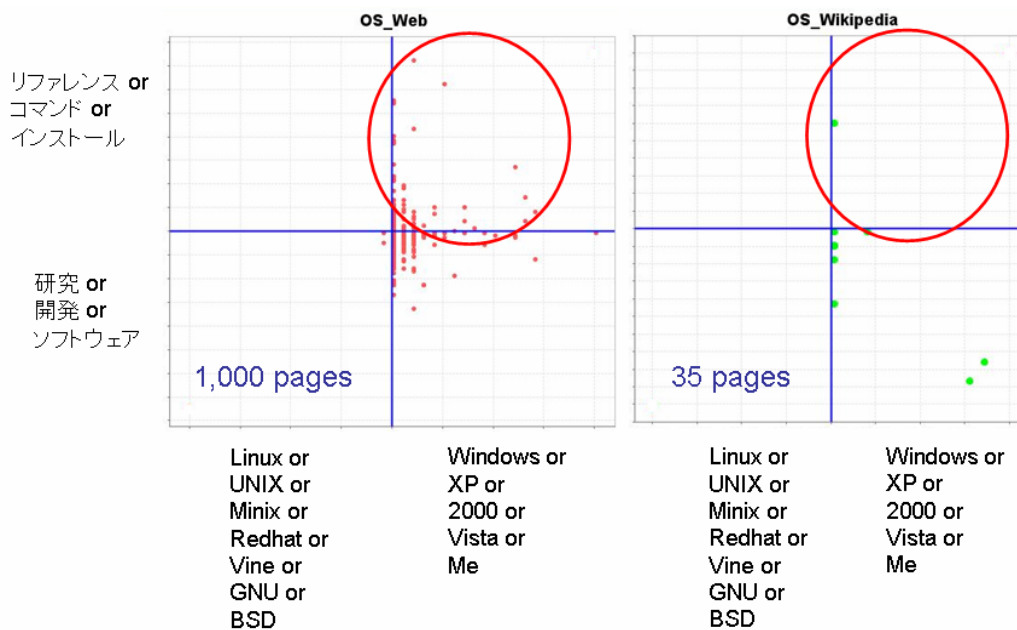


Figure 5.1: Views when topic is “OS”  
(Match-up set: Web, Object set: Wikipedia)

A left view is compared with a right view. Then, the difference can be discovered in the part that enclosed it in a red line. The Web page hardly exists in the part enclosed in a red line of a right view in Figure 5.1 though the Web page exists in the part enclosed in a red line of a left view. The knowledge described on the Web page that exists in the part can be discovered from that and it be discovered that it is insufficient in Wikipedia. That is, it can be discovered that the description that relates to the research, development, the academic society, and the workshop concerning OS is insufficient in Wikipedia.

## 5.2 Application Example 2: “Google”

In Figure 5.2, the topic is a view of “Google”. A left view is a match-up set, and the collection object is Web. A right view is an object set, and the collection object is Wikipedia. With a positive technology or development or retrieval or algorithm or ..direction.. engine the word used for the axis of a horizontal axis. Financial affairs or corporate or stock prices or corporate culture or employee is used for a negative direction of a horizontal axis. Web service or image or news or directory is used for a positive direction of the spindle. Application or desktop or Picasa toolbar or is used for a negative direction of the spindle.



Figure5.2: Views when topic is “Google”  
(Match-up set: Web, Object set: Wikipedia)

A left view is compared with a right view. Then, the difference can be discovered in the part that enclosed it in a red line. The Web page hardly exists in the part enclosed in a red line of a right view in Figure 5.2 though the Web page exists in the part enclosed in a red line of a left view. The knowledge described on the Web page that exists in the part can be discovered from that and it be discovered that it is insufficient in Wikipedia. That is, it can be discovered that the description that relates to an enterprise, financial affairs, and a local application concerning Google is insufficient in Wikipedia.

### 5.3 Application Example 2: “Agent $\cap$ AI”

In Figure 5.3, the topic is a view of “Agent  $\cap$  AI”. A left view is a match-up set, and the collection object is Web. A right view is an object set, and the collection object is Wikipedia. The word used for the axis uses Research or Development or Software or Application for a positive direction of a horizontal axis. University or Laboratory or College is used for a negative direction of a horizontal axis. Conference or Workshop or Meeting is used for a positive direction of the spindle. Thesis or Literature or Book or Text or Textbook is used for a negative direction of the spindle.

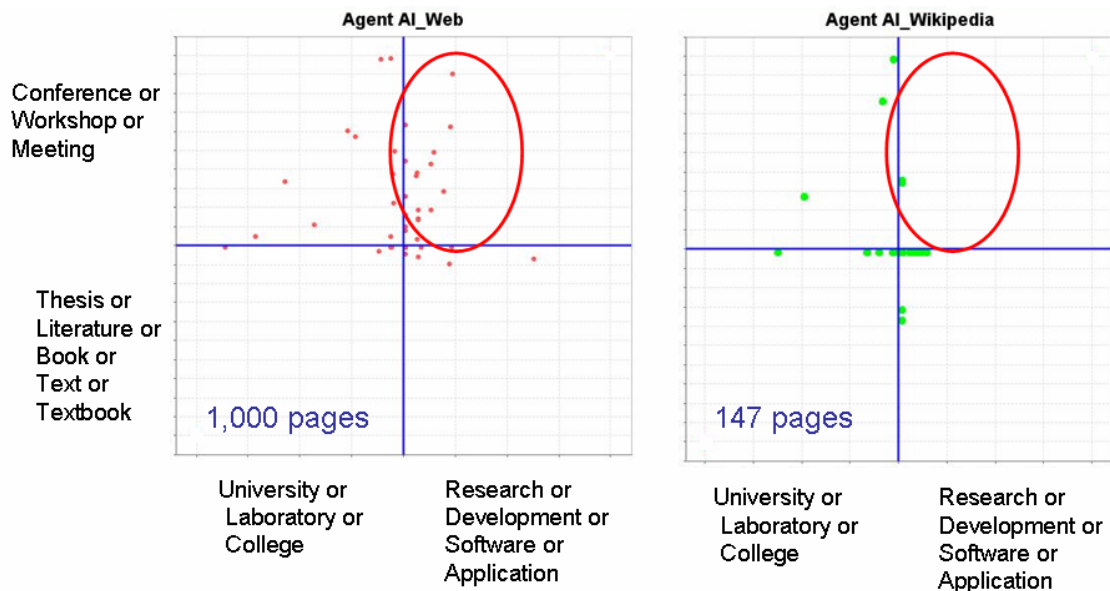


Figure5.3: Views when topic is “Agent  $\cap$  AI”  
(Match-up set: Web, Object set: Wikipedia)

A left view is compared with a right view. Then, the difference can be discovered in the part that enclosed it in a red line. The Web page hardly exists in the part enclosed in a red line of a right view in Figure 5.3 though the Web page exists in the part enclosed in a red line of a left view. The knowledge described on the Web page that exists in the part can be discovered from that and it be discovered that it is insufficient in Wikipedia. That is, it can be discovered that the description that relates to the usage like the reference concerning Agent  $\cap$  AI etc. is insufficient in Wikipedia.

## **Chapter 6 Discussion**

It aimed to support Wikipedia in this research so that the user might easily participate in service that used start Wisdom of Crowds. Concretely, the discovery of the grasp of the whole image of Web page group that related to the topic of Wisdom of Crowds and contents to be able to contribute was assisted by using the visible making the Web page system named PlainView.

The result user who assisted describes what advantage you coming obtain in this chapter based on the application result to Wikipedia in Chapter 5. In addition, how to do so that the user may search efficiently to Wisdom of Crowds is described.

### **6.1 Support for User when Wisdom of Crowds is Formed**

Wisdom of Crowds is for various users to bring together each had knowledge and experience, and to generate information to be worthy as one huge knowledge set. At this time, the user can distinguish when information on Wisdom of Crowds is consumed and when information on Wisdom of Crowds is generated and refined by he or she. The user of the former has the purpose that it wants to examine a clear generally arbitrary topic. Therefore, the item that relates to the topic by using an existing search technique can be efficiently obtained.

However, it is difficult in an existing retrieval to discover information for which information a latter user to be able to contribute and the contribution are needed. Information that oneself can contribute is the one expressed by not a single item but the set. Moreover, when it uses an existing retrieval, the thing whether the contribution is needed should confirm the item one by one. As for the confirmation of all the knowledge that oneself possesses and information of experience by the retrieval, the load added to the user is large. Especially, the scale of Wisdom of Crowds is being made huge now, and it is necessary to do some assistance so that the user may newly contribute from the situation in which the number of items keeps increasing to Wisdom of Crowds.

The grasp of the whole image of Wisdom of Crowds and contents to be able

to contribute were discovered as an assistance method in this research. What advantage you caused for a concrete method of the reason and assistance for which each assistance is necessary and the result users who assisted is described as follows.

### **6.1.1 Grasp of Whole Image of Wisdom of Crowds**

As for contents, the grouped thing of each category is abundant in Wisdom of Crowds as well as structurizing a past Web site. For instance, the item of multi agent is included in agent's category when seeing as an example of Wikipedia, and agent's category and item are included in the category of artificial intelligence.

However, neither user's knowledge nor the set of interest object are necessarily corresponding to grouping performed by Wisdom of Crowds. Therefore, the whole image of Web page group composed of user's knowledge and the viewpoint of interest is requested to be presented.

The user is offering the whole image from user's knowledge and the viewpoint of interest by inputting the word that uses PlainView for a topic and an arbitrary axis. Especially, the same result was only presented to various users in an existing retrieval that used only the topic used in general. The word that uses the same topic for various axes in general is input, and on the other hand, it straightens, and the point that the whole image that adjusts to various knowledge and the interests of various users can be offered is a feature in the method of using PlainView.

### **6.1.2 Discovery of Contents to be able to Contribute**

The user discovers contents to be able to contribute by comparing the views made with PlainView between two or more set. When Web is applied in the match-up set as an example and Wikipedia is applied to the object set, it thinks. At this time, the user collects Web pages by using the same topic for two set. In general, because the number of elements uses the case where it is few from the match-up set for the object set, the collected numbers of pages may be different in two set. It is important how to distribute each Web page in the view to discover contents to be able to contribute, and the number is not important.

Next, if the collection of pages was able to be completed, the view of each set

is made. At this time, the word used for two set is assumed to be the common one though the user of the word used for the axis is arbitrary. The difference between set can be discovered by doing the weight putting in a common word to the Web page of each set. It is because of generation for which a similar view because the view is made as for this as the knowledge of two set is similar by the same weight putting.

That is, the user compares two views, discovers the part where close of the view and sparse part are different, and when the knowledge that should be plotted in the part is insufficient it, the set that relatively has a sparse part is surmisable. For instance, when pages concerning the scientific term are collected in Wikipedia, it can be discovered that information like the academic society, the thesis, the book, and the research base (university, research laboratories or researcher, etc.), etc. is relatively little. The user comes to be able to refine Wisdom of Crowds by discovering, and catching such insufficient information.

The formation of Wisdom of Crowds where be able the balance or more becomes possible by comparing it from various viewpoints when the views are compared though as mentioned above is done and Wisdom of Crowds is refined by the participation of the various users. Concretely, it is preferable that the word used for the axis should not be fixed by the one input once, and changes variously by the interaction with the user. In that case, close of the view and sparse part change because it obtains a different view when a different word is used for the same topic even if the match-up set is similar distribution in a certain word to the object set, too. The user can form the knowledge set where be able the balance by catching knowledge for the same topic to a sparse part in this.

In the same way, various words are applied to the same topic, the user offers knowledge in each view, and Wisdom of Crowds (Compared with the match-up set) becomes possible the formation of the knowledge set without the excess and deficiency.

## **6.2 Efficient Discovery Technique**

The load that the technique used by this research puts on the user must be a minimum because it supports the contribution of the user to Wisdom of Crowds. It is necessary to establish the method of efficiently using the system for that. Especially, the system that uses it by this research comes to compel inefficiency work to the user oppositely when it makes a mistake in use while can satisfied the each user's individual demand. This chapter considers use that enables an efficient search.

It concretely considers it from the side of interaction between the word and the system that uses it for the axis as follows.

### **6.2.1 Words used for Axis**

As for the word used for the axis, weight is added to the Web page when agreeing to the word extracted as a feature word on each Web page. Therefore, when the proper noun is used, big weight is often added. The plot point on the Web page can be expected to be distributed by adding the proper noun that relates to the word used for the axis at that time by OR when the Web page concentrates on one place and the same line when the view is made from this for instance. For instance, it thinks about the case to use Linux and word Windows for the topic of OS. At this time, obtaining the view to which the plot point on the Web page is distributed becomes possible by adding a more concrete proper noun named RedHat·Vine·Gentoo and XP·2000·Vista·Me to these words.

It is possible to add the synonym of the word used for the axis as a similar method. It is a purpose to distribute the plot point on the Web page by increasing a number of words corresponding just like the method of this adding the above-mentioned proper noun, too. For instance, it thinks about the case to use the axis like the research, development, and use, etc. for the topic of OS. At this time, it comes to add application and software <--> commands thought to be a synonym of these words and words of installation. It differs from the synonym in a dictionary meaning a little with the synonym described here. The user uses the word with the relation that I want to give weight as a synonym.

Finally, when the synonym is compared with the case where it adds it, using

the proper noun greatly distributes the proper noun. Therefore, the user comes to be able to distribute it efficiently by using the above-mentioned two methods properly after it thinks how much present view you want to distribute.

### **6.2.2 Interaction with System**

A target view is not necessarily obtained for an operation once though the user inputs the topic and information of axis to the system and makes the view. Especially, PlainView used by this research has the possibility being generated for the view that the user doesn't intend according to the method of setting the word while can satisfied the demand of the user that flexibility is high, and various as an arbitrary word can be set to the axis.

Then, the user comes to be able to inspect the generated view, and to add and to edit the operation if necessary. For instance, can the injury it with weight to consider the synonym and the spoken language by changing the distribution of the view by changing the word used for the axis, and adding the word becomes, and the range of distribution can be enlarged. Figure 6.1 is a sequence chart where the interaction of the system is shown the user.

Moreover, if the views are compared from various viewpoints by changing the word used for the axis many times, refining Wisdom of Crowds to the knowledge set where be able the balance becomes possible.

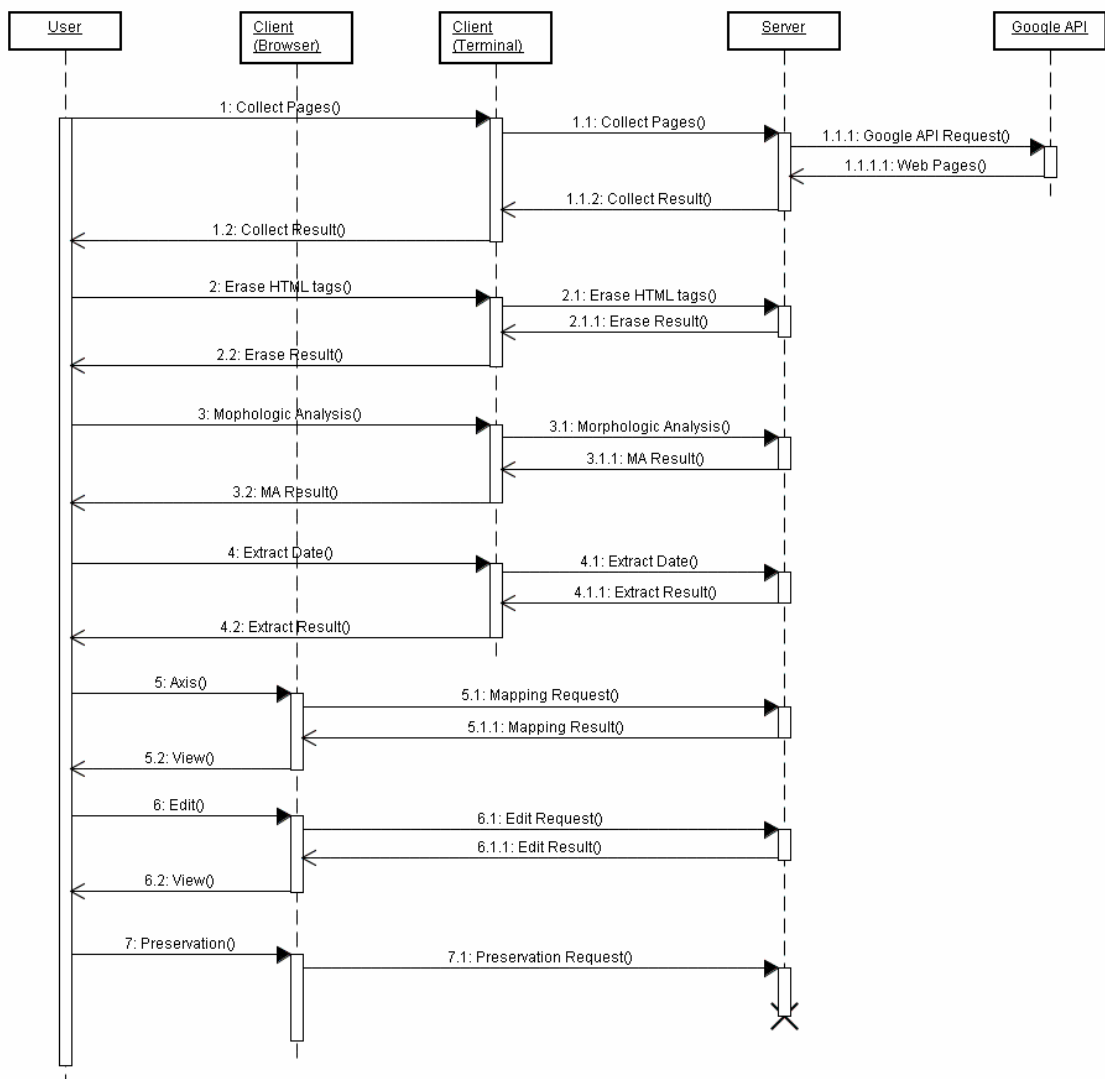


Figure 6.1: Sequence chart of making view

### 6.3 Use of Wisdom of Crowds

The user came to be able to acquire information wanting it easily by developing and generalizing Web. All of kinds of information overflowed on Web today, and the user came to be able to obtain various knowledge there though obtained information was limited at time when information on Web was scarce. There are various information etc. that a general second information where official information and the mass media that the enterprise, the university, and government and municipal offices had disclosed on Web as

one means of media described a report, a criticism to those official information, and an original point under discussion, user is not basically owed the obligation to information to information and made remarks on on Web of various topics. As for them, volume of information is little though the reliability of official information is high. There is a feature of variously obtaining a large amount of information though reliability is low for various information.

Media literacy as the ability to understand reliability and the purpose of information has come to be requested from such a present situation by the user who uses Web. Moreover, those attainments not only are requested from the user but also the attempt that starts measuring the systematization of reliability and information when seeing from the side of the system named Web is performed. However, a lot of time is necessary for problems for solution by the time they spread widely as an infrastructure and the purpose is achieved actually.

Such inside appearance is thought to be one effective method for the current of use of done Wisdom of Crowds to systematize knowledge. In Web now at the time of exist together information described from a viewpoint different because of a different format, the ability for that is necessary for discovering information assumed to be user's purpose from the average adequately and to read it through because it is described by a different form, discovered information needs user's labor. As for the Web page of described Wisdom of Crowds, it formats to them the same, and the labor needed when target information is discovered and it is read through is fewer than general Web from the same viewpoint.

Referring to the item can obtain knowledge efficiently when it is examined whether the item is in Wikipedia before it retrieves it with Web when the examination thing is actually done, and exists. The item of Wikipedia is always added, is being refined by various all over the world, users now, and has enhanced value as the knowledge data base systematically arranged. And, the user comes to be able to obtain various knowledge up to now more efficiently by using the knowledge data base.

## Chapter 7 Conclusion

It paid attention to the trend of activation Web of the use of Wisdom of Crowds in this research, and it worked on the solution of the problem when Wisdom of Crowds was formed. The user comes to be able to share knowledge and the experience by using Wisdom of Crowds. A lot of various information except user's purpose exist in Web though Web can be considered to be a set of knowledge in a broad sense. In Wisdom of Crowds , for example, Wikipedia, the point to settle information that becomes such a noise from can every edit for the item described once to one refined item is different.

Moreover, the difference is seen in the use etc. of the amount and the term in Web in the form of the description. Therefore, the load for which information described by the composition in which the system is relatively set up like Wikipedia by the same format is more necessary for the user's understanding can be reduced.

There is a problem when each item describes knowledge and the experience it is necessary to set up the system as a whole though there is an advantage of it not is in current Web in Wisdom of Crowds like this. For instance, when contributing to Wisdom of Crowds based on knowledge and the experience, the user should understand an insufficient part for Wisdom of Crowds by information that he can offer. However, the means to access Wisdom of Crowds at present is only an existing retrieval, and the user should image the whole image of Wisdom of Crowds by inspecting the item obtained by the retrieval one by one.

Then, it thought it paid attention to the point "Offer of various views" that was the feature of Plain View in this research, the feature was made the best use of, and the formation of Wisdom of Crowds was supported. The whole image of Wisdom of Crowds was presented to the user by PlainView, and the discovery of the item for which the contribution was needed was facilitated. As a result, it is supported to participate in the formation of Wisdom of Crowds efficiently. Moreover, the evaluation and consideration were done by using the concept of Complementary search in that case. The evaluation method where it

shows by this research is different from the evaluation method like a past recall ratio and the relevance ratio, etc. Concretely, it is evaluated to bring the whole image of Wisdom of Crowds of the object set in the definition of Complementary search close to the whole image of Wisdom of Crowds of the match-up set aiming. The user can refine Wisdom of Crowds by comparing the match-up set with the view of the object set, and supplementing insufficient knowledge for the object set. In addition, those proposal techniques were actually applied to Wikipedia and it evaluated it. As a result, it was able to be confirmed to a related item of the scientific term to be able to discover that the academic society and thesis information were little in Wikipedia. Moreover, it has been understood to obtain a target view easily when using it for the word used for the axis while considering the proper noun and the synonym. And, it has been understood to be able to form Wisdom of Crowds where be able the balance compared with the match-up set by comparing the views from various viewpoints by generating the view interactively to the system, and adding of the item and editing it.

It is clear that the use of Wisdom of Crowds, for example, Wikipedia actively becomes it more and more as a means to share knowledge and the experience in the future. In that case, it is thought that the load when the user contributes to Wisdom of Crowds can be decreased by this research.

## **Acknowledgments**

I would like to express my sincere gratitude to Professor Toru Ishida at Graduate School of Informatics, Kyoto University, for invaluable advice and discussion.

I am also grateful to my research adviser, Professor Katsumi Tanaka at Department of Social Informatics, Graduate School of Informatics, Kyoto University, Professor Hirokazu Tatano at Department of Social Informatics, Graduate School of Informatics, Kyoto University, and Associate Professor Mizuho Iwaihara at Department of Social Informatics, Graduate School of Informatics, Kyoto University.

I would like express special thanks all the members of Ishida Laboratory at Kyoto University for their support.

## References

- [1] Dave, K., Lawrence, S. and Pennock, D. M.: Mining the peanut gallery: opinion extraction and semantic classification of product reviews, WWW'03: Proceedings of the 12th international conference on World Wide Web, New York, NY, USA, ACM Press, pp. 519–528 (2003).
- [2] Hu, M. and Liu, B.: Mining and summarizing customer reviews, KDD'04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM Press, pp. 168–177 (2004).
- [3] Liu, B., Hu, M. and Cheng, J.: Opinion observer: analyzing and comparing opinions on the Web, WWW '05: Proceedings of the 14th international 52 conference on World Wide Web, New York, NY, USA, ACM Press, pp. 342–351 (2005).
- [4] Tateishi, K., Ishiguro, Y. and Fukushima, T.: A Reputation Search Engine That Collects People's Opinions Using Information Extraction Technology, IPSJ Transaction on Databases, Vol. 22, pp. 115–123 (2004).
- [5] Daisuke Nakamura: User-Centered Approach to Visualizing Web Pages, Master thesis on Department of Social Informatics, Graduate School of Informatics, Kyoto University (2006).
- [6] Plaisant, C., Mushlin, R., Snyder, A., Li, J., Heller, D. and Shneiderman, B.: LifeLines: Using Visualization to Enhance Navigation and Analysis of Patient Records, Technical Report CS-TR-3943 (1998).
- [7] Rekimoto, J. and Green, M.: The Information Cube: Using Transparency in 3D Information Visualization, WITS '93: Proceedings of the Third Annual Workshop on Information Technologies & Systems (1993).
- [8] Lamping, J., Rao, R. and Pirolli, P.: A focus+context technique based on hyperbolic geometry for visualizing large hierarchies, CHI '95: Proceedings of the SIGCHI conference on Human factors in computing systems, New York, NY, USA, ACM Press/Addison-Wesley Publishing Co., pp. 401–408 (1995).
- [9] Munzner, T.: H3: Laying out large directed graphs in 3D hyperbolic space, Proceedings of the 1997 IEEE Symposium on Information Visualization,

- pp. 2–10 (1997).
- [10] Hearst, M. A.: TileBars: Visualization of Term Distribution Information in Full Text Information Access, CHI '95: Proceedings of the SIGCHI conference on Human factors in computing systems (1995).
  - [11] Hearst, M. A. and Karadi, C.: Cat-a-Cone: an interactive interface for specifying searches and viewing retrieval results using a large category hierarchy, Proceedings of SIGIR-97, 20th ACM International Conference on Research and Development in Information Retrieval, pp. 246–255 (1997).
  - [12] Lagus, K., Kaski, S. and Kohonen, T.: Mining massive document collections by the WEBSOM method, Inf. Sci., Vol. 163, No. 1-3, pp. 135–156 (2004).
  - [13] Rao, R. and Card, S. K.: The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information, CHI '94: Proceedings of the SIGCHI conference on Human factors in computing systems, New York, NY, USA, ACM Press, pp. 318–322 (1994).
  - [14] Shiozawa, H., Nishiyama, H. and Matsushita, Y.: The Natto View: An Architecture for Interactive Information Visualization, Transactions of IPSJ , Vol. 38, No. 11, pp. 2231–2342 (1997).
  - [15] Shneiderman, B.: Designing the User Interface, Addison-Wesley (1992).
  - [16] Mackinlay, J. D., Robertson, G. G. and Card, S. K.: The Perspective Wall: Detail and Context Smoothly Integrated, CHI '91: Proceedings of the SIGCHI conference on Human factors in computing systems, Addison-Wesley, pp. 173–179 (1991).
  - [17] Robertson, G. G., Mackinlay, J. D. and Card, S. K.: Cone Trees: animated 3D visualizations of hierarchical information, CHI '91: Proceedings of the SIGCHI conference on Human factors in computing systems, ACM Press, pp. 189–194 (1991).
  - [18] Q Ma, A Nadamoto, K Tanaka: Complementary Information Retrieval for Cross-Media News Content: Proceedings of the ACM International Workshop on Multimedia Databases, ACM Press, pp. 45–54 (2004).