How Does a User Utilize a Chart-based Interface to Conduct Exploratory Data Analysis?

Kazuhiro Tanaka
Kansai University
2-1-1 Ryozenji, Takatsuki
Osaka, 569-1095 JAPAN
k316176@edu.kutc.kansai-u.ac.jp

Daiki Hasui
Kansai University
2-1-1 Ryozenji, Takatsuki
Osaka, 569-1095 JAPAN
fa80357@edu.kutc.kansai-u.ac.jp

Mitsunori Matsushita
Kansai University
2-1-1 Ryozenji, Takatsuki
Osaka, 569-1095 JAPAN
mat@res.kutc.kansai-u.ac.jp

ABSTRACT

This paper proposes an information retrieval system that utilizes a chart-based interface to support a user's exploratory data analysis. In such analysis, a user tends to examine data to support his/her hypothesis and to seek new perspectives about the data. To conduct this analysis, a user usually overviews data from various viewpoints first, compares multiple data to find differences and similarities among the data, and then accesses the original texts (e.g., newspaper articles) that contain the data. Our proposed system aims to support such process by providing a chart-based interface, which displays two charts side by side and permits a user to overlay the two charts for comparison, and then accesses the original texts. This paper presents a basic framework of the target task (VisEx), requirements for the task support system, and the details of our proposed system. The result of the users' study conducted in VisEx task are reported after analyzing the obtained data, which includes log records and post-questionnaire responses.

Categories and Subject Descriptors

H3.3 [Information Storage and Retrieval]: Information Search and Retrieval - search process, selection process

General Terms

Human Factors

Keywords

Mulitiple Charts, Exploratory Data Analysis, Information Visualization, Trend Summarization

Team Name

KUTC

Subtasks/Languages

Japanese Trend Collection

External Resources Used

MuST Corpus in NTCIR-7

1. INTRODUCTION

Large amounts of data have accumulated on the Internet in recent years. Various types of data are available, for instance, numerical data sets, such as those describing electricity consumption, and text data, such as newspaper articles. Such data are important as not only an archive of real-world events, but also a resource for obtaining useful knowledge, observing trends, and making novel findings to achieve rational decision-making and problem-solving. When analyzing this data, an analyst repeats a trial-and-error search process and examines the retrieved data in order to find support for his/her hypothesis and seek new perspectives.

The goal of our research is to develop a system that supports such a user's information-seeking activity, which is called exploratory data analysis [1]. Exploratory data analysis is an iterative process that consists of several different steps. This paper focuses on the following three steps in this process: (1) The analyst overviews data from various perspectives in order to grasp overall trends (data overview), (2) The analyst compares data to find characteristic patterns, differences, and similarities (data comparison), and (3) The analyst accesses detailed information such as the original texts (e.g., newspaper articles) that contain the data in order to deepen his/her knowledge (detailed information access).

This paper proposes an information retrieval system in which a chart-based interface supports exploratory data analysis. The system displays two charts side by side and permits a user to overlay them for comparison and then access more detailed information. In addition, the system enables users to easily repeat a trial-and-error series by linking the functions seamlessly. This paper reports the results of a study of user behavior conducted in the Interactive Visual Exploration task (VisEx); the study analyzes the obtained data, which includes log records and post-questionnaire responses.

2. TASK OVERVIEW

2.1 VisEx Task

VisEx[1] is a pilot task of NTCIR-9. Its purpose is to establish a framework for evaluating explorative information access environments, in which a user interactively refines and/or elaborates his/her information needs. By repeating various activities, including query articulation, search, and reflection, the user collects the appropriate information. In such environments, information visualization techniques are important for displaying access results and providing visual keys to how to interact with the data resources. VisEx aims to evaluate such information access environments in a comprehensive manner.

[1] http://must.c.u-tokyo.ac.jp/visex/

The organizers of VisEx provide a common framework for explorative information access environment systems (IAESs) and design experimental tasks. VisEx participants[2] submit the core of an IAES, which works in the common framework. Submitted IAESs are evaluated through laboratory experiments with human subjects, who are asked to complete the experimental tasks in the given environments. Objective data such as the elapsed time and number of interactions are measured, and subjective data such as the (dis)satisfaction of the subjects are also examined through questionnaires.

2.2 subtasks

VisEx in NTCIR-9 employs two types of task: the Event Collection Task and the Trend Summarization Task. The details of each subtask are as follows.

- Event Collection Task

  The Event Collection Task uses the event-list questions in the NTCIR-7 ACLIA Task as the test set, which presents questions as requests for needed information. For example, a participant may ask "Please tell me about incidents where NATO has recognized cases of friendly fire" or "Please tell me about airplane crashes that have happened in Asia." This task requires subjects to collect as many nuggets as possible in a given time period, where the nuggets are event characteristics such as time and place.

- Trend Summarization Task

  The Trend Summarization Task requires that subjects summarize the trends in time-series statistical information, such as the subjects of NTCIR-5, 6, and 7 Multimodal Summarization for Trend Information (MuST)[3]. Subjects are asked to summarize not only changes in a given statistic but also the background and influences. For example, the needed information can be expressed as "Please give me a summary of the state of the cabinet approval rating from 1998 to 2001" or "Please give me a summary of the changes in gasoline price from 1998 to 2001." In this task, subjects are asked to collect as many nuggets as possible in a given time period, where the nuggets are basic information that constitutes the requested summary.

In both tasks, in addition to the functionality that is required for a user's successful exploration, it is important for the IEAS to be a comfortable environment with little stress.

3. CHART-BASED INTERFACE

3.1 System design

Among the processes that a user executes in an exploratory data analysis, comparison is one of the chief activities. The heart of many information visualization and data mining techniques is a single question: "Compared to what? [8]" Therefore, we focus on comparison. We think visualization via a chart is effective when users consider this question.

[2]Note that a team that submits an IAES for a VisEx task is referred to as a participant, whereas people who participated in the user study conducted in VisEx are referred as subjects.
[3]http://must.c.u-tokyo.ac.jp/

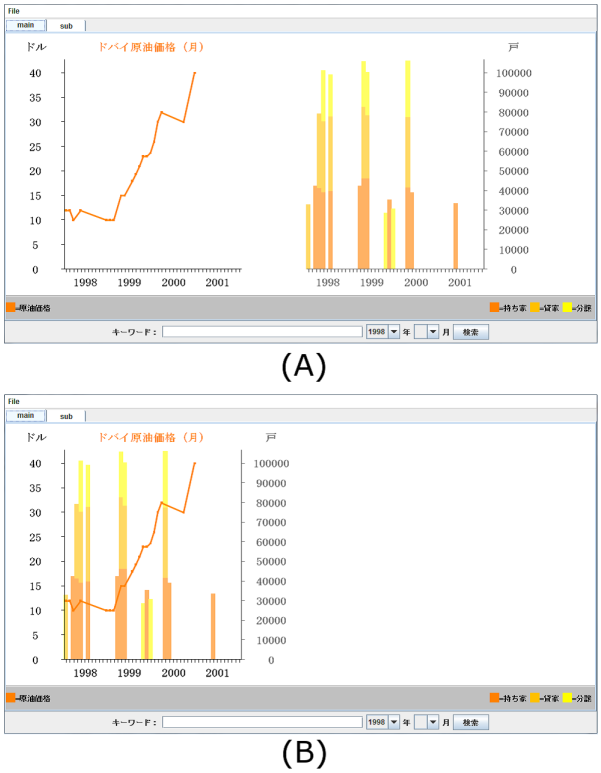

Figure 1: Images of proposed chart-based interface. (A) Placing charts side by side, (B) Overlaying charts

Visualization on a chart is suitable for surveying trends in the data and enables easy comparison of the data in order to discover correlations between several sets of data.

In the field of HCI, it is said that a useful starting point for designing an advanced graphical user interface is the Shneiderman's Visual Information Seeking Mantra [5]: "Overview first, zoom and filter, then details-on-demand." In an exploratory data analysis, analysts investigate data without having a clearly stated goal of how to use the data in advance. Therefore, they first establish their hypothesis by overviewing data from various perspectives.

On the basis of the hypothesis, they compare the data in order to find differences and similarities among them and extract the necessary data from a large amount of data. That is, they seek evidence that affirms or denies their hypothesis. If they find interesting data within a large amount of data or want to inspect their hypothesis in detail, they examine the data more deeply. If they reject the hypothesis, they seek a new perspective on the analysis. This process is not a one-time interaction but is repeated many times.

We aim to implement a system in which users can easily repeat this process: overview data from various perspectives, compare them, and access detailed information as desired.

3.2 Implementation

Based on the system design described in section 3.1, we implemented an information retrieval system as a prototype. The prototype system enables a user to retrieve the common data sets in VisEx. The document sets consist of Mainichi newspapers in Japanese from 1998 to 2001.

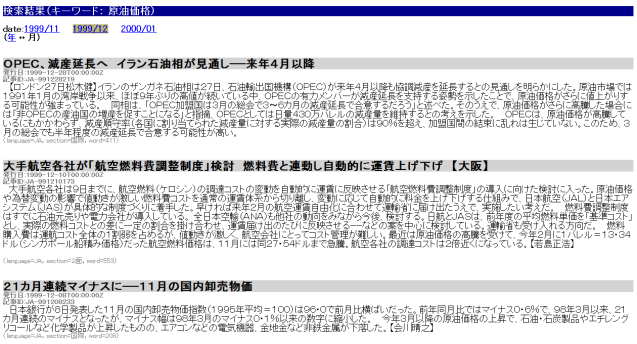

Figure 2: Search result example

Before the experiments, we manually prepared data set for presentation on a chart. These data came from the MuST corpus [3], a data set containing tagged elements for visualization of statistics. Hereafter, the generated data is referred as plot data. A plot data is generated by extracting necessary elements to visualize such as a topic name, a statistic, and a unit on the basis of the statistic's name, from the MuST corpus. We also attached a search word to each plot data. The search word was obtained basically from the topic name of MuST corpus. The attached search word on each plot data can be used for judging whether the plot data is suitable to the user's exploration content or not when a user submits a generic keyword.

We gathered information from a time series of articles about 27 topics, such as gasoline prices and the number of computers shipped, covered in the Mainichi newspaper from 1998 to 2001. From the gathered information, we prepared 67 pieces of plot data, from which users select data to visualize. The plot data are listed under the topic name. Users can visualize the data by selecting a topic name from the list. We define running their eyes over the list as overviewing the data, although the list represents only a portion of the common data sets.

Figure 1 shows screenshots of the interface. In Figure 1-(a), two charts are displayed side by side. Visualizations take the form of line charts or stacked bar charts, depending on the type of data. The interface enables a user to compare two pieces of data using these charts. Figure 1-(b) shows an overlay of two charts in the interface. The interface permits users to move the charts using a mouse. Therefore, it enables users to compare data in a more exploratory manner in order to find correlations.

The interface enables a user to retrieve data by clicking a data point plotted on a chart. More detailed information (newspaper articles) is retrieved according to the point the user clicks and the search word in the plot data.

Figure 2 shows an example of a search result. In this example, the user clicked a date "December 1999", and the topic name "Dubai crude oil price." The query is "crude oil price," and the system searches articles that were published in December 1999. By such interactions, the system makes it easy to access background information from the charts. In addition, the upper part of the search result screen permits a user to change the date of the retrieved publications to the following month (year) or previous month (year), and to change the level of granularity (month or year). As a result,

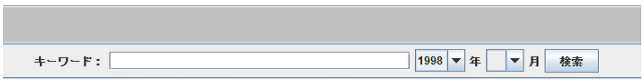

Figure 3: Keyword input box

the user can read the search results along a time series and at different levels of granularity.

In addition, the interface provides a keyword search. As shown in Figure 3, the lower part of the interface contains an input box, a pull-down menu for selecting of a time period, and a search button. Users enter requests in the input box and use the pull-down menu to select a date from "January 1998" to "December 2001". The keyword search allows the user to search when the plot data does not contain information that the user wants to see.

4. EXPERIMENT

4.1 Experimental design

We conducted an experiment using the Trend Summarization subtask in VisEx. Subjects were asked to summarize the trends in time-series statistical information on a given topic by collecting not only the values and variations in the statistics but also the reasons for the changes and pertinent influences. We used the document sets in VisEx described in section 3.2.

The organizer of VisEx provides a baseline system. Its interface design is simple and similar to that of ordinary web search engines. When a user enters keywords into a query form, the retrieved results are returned in an ordinary list format. Although an AND search is performed by default, OR and phrase searches are also available. The search results can be ordered in descending/ascending order according to the score or release date. Each entry in a search result consists of an article's release date, ID, and title and a snippet. Clicking the title displays the entire contents of the article. This system was also used in training sessions so that users could understand the task and to provide a reference for comparison with the submitted system.

The VisEx experiments were conducted from Aug. 15-18, 2011. The experiment using our system (the KN System) was conducted on Aug. 16, 2011, and that of the baseline system (the BL System) was conducted on Aug. 18, 2011. A between-subjects experimental design was employed; i.e., different users used the KN System and the BL System. The number of users for each system was five (male: two, female: three).

In the experiment, each subject was requested to conduct an exploratory search on the given four topics; the time limit for each experiment was 50 minutes. The topics used in the experiments are described below.

T1 Please examine the situation regarding gasoline. I would like to know about changes in Dubai crude oil prices and regular gasoline prices.

T2 Please examine the evaluation of the Cabinet. I would like to know about changes in the approval and disapproval ratings.

T3 Please examine the employment situation. I would like

to know about changes in the number of unemployed people and the unemployment rate.

T4 Please examine the demographic composition of Japan. I would like to know about changes in the elderly population, young population, and birth rate.

The experimental data consist of reports made by the subjects as direct products of the tasks, log records of subjects' information access behaviors, and subjective evaluations collected using questionnaires completed by the subjects.

4.2 Experimental result

Table 1 shows the experimental results for the subjects that used the BL System (F18, F19, F20, M19, and M20). Table 2 shows the experimental results for the subjects that used the KN System (F06, F07, F08, M06, and M07). In Table 2, "Keyword," "Clicking a point," "Opening data," "Moving a chart," and "Changing date" indicate the number of times a keyword search was used, data were retrieved by clicking a point on a chart, a list of plot data was opened, a chart was moved (including overlaid charts), and the publication date and level of granularity were changed, respectively. Table 3 shows the results of a questionnaire that asked users how useful, functional, and efficient they found the target system, and whether they want to use it again, and want to recommend it to others. The questionnaires used a seven-point scale.

Our proposed system, the KN System, aims to support a user's trial-and-error activity in exploratory analysis, in which a user investigates data without a definite predetermined purpose. In this experiment, however, subjects were not asked to find a fact something interesting under his/her own concern, but asked to make a report on the given topics. This implies that it is difficult to evaluate whether our impremented system is suit for the design direction or not.

So, we assumed that ethnomethodological observation of a user's activity in each step, mentioned in section 1, is the objective of the analysis, that is, we observe the subjects' activities under the following three steps: (1) data overview, (2) data comparison, and (3) detailed information access.

In overviewing, subjects surveyed a list of plot data or visualized data. According to Table 2, a list of plot data was opened comparatively frequently in T3 but infrequently in T2 and T4. In T3, the subjects selected mainly plot data ("number of unemployed" and "unemployment rate"). Only F06 visualized other data ("effective opening-to-application ratio," "coincident indicator of economic trend," and others). Therefore, we estimated the plot data in T3 to be rich. Thus, users would likely open a list of plot data frequently.

Multiple charts were compared by placing two charts side by side and overlaying them. According to Table 2, charts were moved comparatively frequently in T3 but infrequently in T2 and T4. Therefore, we conjectured that the frequencies with which lists of plot data were opened and charts were moved may be correlated. This conjecture is supported by the fact that in T1, M06 and M07 also frequently opened data and moved charts. Thus, we suppose that the richness of the plot data they want to visualize influences the frequency of chart moving. In addition, in T1, M06 and M07 reported a negative correlation (Japanese gasoline price fell while crude oil prices rose) in the post-questionnaires. This discovery was not made in the BL System. Considering these results, we suppose that overlaying charts may be effective for finding correlations or negative correlations. However, we think that this point needs further investigation.

In detailed information access, news paper articles were retrieved by clicking a point on a chart or using a keyword search. According to Table 2, users clicked a point on a chart in frequently in T4. In contrast, keyword searching was frequent in T4. We attribute this to the fact that the presented task included the keywords "elderly population" and "young population." Therefore, the user's needs would become clear. In addition, the keyword search was used infrequently in T2 and T3. We consider that the reason for this in T3 is that many plot data were available. On the other hand, this is not the case in T2. We provided the keyword search for cases where there are not many data that users would want to visualize. Despite the lack of plot data, the keyword search was used infrequently because users'search intentions did not become clear.

In all the experiments, the edit time in the KN System tended to be shorter than that in the BL System. In other words, in our system, more time was spent reading newspaper articles and using the system. Further, according to Table 3, the KN System received more positive evaluations than the BL System. Therefore, the result probably implies that the subjects could efficiently produce a summary report using the KN system. We also analyzed log records of the publication date of articles the subjects read when using the KN System. The records showed that they tended to read articles along a time series, and this tendency became stronger in the latter topic. We expected users to pay attention to characteristic points and changes in the charts, but they did not. The reason is believed to be that observing data along a time series enables them to produce a summary report more quickly.

Note that F07's log record can not be examined because she inadvertently used the BL system rather than the KN system in the experiment.

5. DISCUSSION

In the data overview process, we assumed that overviewing and comparing data would take more time when rich data were available in the list of plot data. If the plot data were not rich, the keyword search would be useful when the user's purpose for the analysis is defined to some extent, as it is in this experiment. Here, we manually prepared plot data in advance, but that is impossible in real-world analysis. If the user's purpose is not defined, we think that the system should recommend plot data to the user. Therefore, we need to consider how to establish a standard for recommending plot data from a data set.

In the data comparison process, the information that users found on two charts could reveal a particular comparison, for example, if they discovered a negative correlation. In exploratory analysis, users extract the necessary data from the data set. To support this activity, we believe that users must be able to find not only correlations and negative correlations but also differences and similarities among different data. Therefore, we suggest that a list of plot data should also be visualized on the charts, although we supplied only the topic name for the data in this implementation. In addition, characteristic points on the charts were often ignored, although we expected users to pay attention to them. The reason may be not only that observing along a time series allowed users to produce a summary report more quickly,

but also that when few data were available for comparison, the user compared data having similar names and/or trends, and several pieces of data which didn't have characteristic points. We need to examine the usefulness of data comparison.

In the process of accessing detailed information, users appeared to seamlessly access data, moving from charts to newspaper articles. However, in the post-questionnaire responses, users were not satisfied with the seamlessness of operability after the search results were displayed on a browser. In addition, users felt that it was hard to click points on charts, and the search precision was poor. We need to improve these problems. In particular, we will consider the poor search precision.

Strictly speaking, charts are actual summaries of statistics in a data set for a topic. Therefore, newspaper articles from which the statistics on the charts were extracted should be displayed when users click a data point on a chart. However, in this experiment, the newspaper articles that were displayed were search results prepared manually in advance by searching on a search word. Therefore, we should implement a system in which a chart is a summary of a topic based on data sets such as newspaper articles, and clicking a point on the chart allows user to access the original data.

6. RELATED WORKS

As mentioned in Section 3.1, our proposed interface focuses on supporting a user's "comparison" process conducted in his/her exploratory data analysis. Several systems which take the similar objectives have been proposed.

Takama et al. [7] proposed an interactive visualization system that supports a user's exploration with spatiotemporal trend information by combining statistic charts with geographical map in Japan. This system enables user to compare data sets in terms of area and date. It also permits user to change granularity temporally or spatially.

Matsushita et al. proposed an interactive visualization system named InTREND [4]. The system supports user's iterative exploration by interpreting user's information needs given as a natural language query, and presents a statistical chart as a result of the query. InTREND encourages iterative exploration by maintaining the context of past interactions and utilizing the obtained context to complement the lack of the user's fragmentary query.

These systems utilize a single window though, user can execute comparison activity in his/her exploration because the systems adopt a temporally changing visualization technique. In contrast, there are several systems that utilize multiple window panes for easy comparison between different data sets or different aspects of a data.

Ito et al. proposed a system that employs a hierarchical data visualization technique [2]. The system visualized two different aspects of a data set on separated panes. The system permits a user to manipulate visualized components, so that users can arbitrarily access/explore the detail of the target data. When the user manipulate the data via one visual pane, the visualized content on the other pane also changes.

Stasko et al. proposed an exploration support system named Jigsaw, which supports an analysis with different kinds of investigative and sensemaking scenarios based on textural documents [6]. The system adopts a multi-view system, including a number of different visualizations of the documents in the collection and the entities within those documents.

Our primary purpose is to support a user's comparison between two series of statistical data, which is often observed his/her exploration, in order to discover differences, similarities, or correlations latent in the data. Our proposed system is designed to meet this purpose by placing charts side by side, overlaying them, and accessing more information from them. From this perspectives, the difference between our system and the other systems mentioned above is that the scope of our current system so far is more narrow and, therefore, concrete in terms of the focused user's interaction behavior.

7. CONCLUSION

This paper proposed an information retrieval system in which a chart-based interface supports exploratory data analysis. Our proposed system aims to support a user's iterative exploration activity, which consists of the following three processes; (1) data overview, (2) data comparison, and (3) detailed information access. A prototype system is implemented to meet the user's requirement by focusing on the second process, that is, a "comparison" process. We conducted an experiment using the Trend Summarization subtask in VisEx. As the result, we found that overlaying charts seems to be effective for finding correlations or negative correlations in a comparison.

A lot of future work still remains; we will investigate a novel interface that enables a user to overview more comparable by visualizing plot data. In addition, the interface displays a charts as a trend summarization of data sets and enables a user to access the original data such as newspaper articles seamlessly by clicking a point of a chart.

8. REFERENCES

[1] F. Hartwig and B. E. Dearing. Exploratory Data Analysis. SAGE publications, 1979.

[2] T. Itoh and H. Tachibana. Visualization of corpus data by a dual hierarchical data visualization technique. In Proc. 7th NTCIR Workshop Meeting, pages 534 – 541, 2008.

[3] T. Kato, M. Matsushita, and N. Kando. MuST: a workshop on multimodal summarization for trend information. In Proc. 5th NTCIR Workshop Meeting, pages 556 – 563, 2005.

[4] M. Matsushita. Supporting exploratory data analysis by preserving contexts. In Proc. 9th International Conference on Knowledge-Based Intelligent Information and Engineering Systems, pages 540 – 546, 2005.

[5] B. Shneiderman. Designing the User Interface. Addison Wesley, 2003.

[6] J. Stasko, C. Görg, and Z. Liu. Jigsaw: supporting investigative analysis through interactive visualization. Information Visualization, 7(2):118 – 132, 2008.

[7] Y. Takama and T. Yamada. Visualization cube: Modeling interaction for exploratory data analysis of spatiotemporal trend information. In Proc. 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, pages 1 – 4, 2009.

[8] E. R. Tufte. Envisoning Information. Graphics Press, 1990.

Table 1: Experimental results for the subjects that used the BL System

Topic	Subject	Times of retrieval	Editing time	Ratio of editing time to total time
T1	F18	17	1808	0.59
	F19	11	959	0.495
	F20	5	1297	0.42
	M19	6	2305	0.74
	M20	8	2171	0.7
T2	F18	2	2261	0.77
	F19	14	1189	0.4
	F20	1	1474	0.5
	M19	1	2140	0.73
	M20	12	1839	0.7
T3	F18	13	1321	0.43
	F19	26	991	0.32
	F20	2	1478	0.48
	M19	13	1675	0.52
	M20	8	1689	0.55
T4	F18	29	918	0.29
	F19	24	1215	0.38
	F20	17	679	0.21
	M19	40	1675	0.52
	M20	8	2123	0.66

Table 2: Experimental results for the subjects that used the KN System

Topic	Subject	Retrieval	Edit	Ratio	Keyword	Clicking a point	Opening data	Moving a chart	Changing date
T1	F06	3	1345	0.465	0	3	2	1	12
	F07	26	1436	0.495	1	25	2	1	1
	F08	15	1321	0.452	3	12	2	0	35
	M06	16	1033	0.353	0	16	8	4	62
	M07	14	797	0.274	3	11	7	3	3
T2	F06	10	506	0.17	8	2	3	2	5
	F07	26	1375	0.462	0	26	1	0	0
	F08	14	1564	0.523	0	15	1	0	17
	M06	10	1227	0.41	0	10	3	0	35
	M07	25	734	0.248	0	25	1	0	13
T3	F06	8	1197	0.388	2	6	9	8	7
	F07	36	842	0.271	0	36	2	7	0
	F08	10	1423	0.455	0	10	3	0	12
	M06	6	817	0.262	1	5	6	5	52
	M07	21	578	0.185	0	21	6	8	4
T4	F06	14	1144	0.398	14	0	2	1	2
	F07	==	==	==	==	==	==	==	==
	F08	15	1269	0.439	8	7	3	0	43
	M06	5	994	0.344	4	1	2	0	33
	M07	24	815	0.28	18	6	2	1	64

Table 3: Results of a post-questionnaire of the BL system and the KN system

	KN						BL					
	F06	F07	F08	M06	M07	Avg.	F18	F19	F20	M19	M20	Avg.
Useful	5	5	4	7	6	5.4	2	4	6	4	4	4.0
Functional	6	3	5	4	4	4.4	2	4	5	2	4	3.4
Efficient	6	2	4	5	6	4.6	2	4	5	2	3	3.2
Want to use again	6	2	4	5	5	4.4	2	3	4	1	3	2.6
Want to recommend	5	2	4	4	6	4.2	2	3	3	1	2	2.2