Rutgers Information Retrieval Evaluation Project on IR Performance on Different Precision Levels

럿거스 정보검색 평가 프로젝트에 관한 연구

Hyuk-Jin Lee*, Nicholas J. Belkin, Bob Krovitz

ABSTRACT

The purpose of this study is to investigate what level of difference in precision would be significantly perceived by a human user of an information retrieval system. Not many researches have been conducted with regards to this issue in information retrieval field. Despite the non-significant results, there were several interesting findings in recognizing different levels of precision rates. The correctness of relevance task had little to do with the taken time for the task. In addition, the strong relationship between the subjects' topic familiarity and rate of correct judgments is one of the most interesting results in this study. It turned out that the subjects have more difficulty in a situation they have to judge between the two lists having more non-relevant documents than in a situation they do between the lists having more relevant documents. Finally, the serious influence from the first top N documents in a list for relevance judgment task has been confirmed.

초 록

이 논문의 주요목적은 정보이용자들이 어떤 수준의 정확률 차이에서 유의미하게 차이를 인지하는지를 알아보고자 하는 것이다. 그에 관련한 몇 가지 흥미 있는 결과가 도출되었다. 그 외에 적합성 판정은 이용자의 판정시간과 관계가 없는 것으로 나타났다. 그리고 주제에 대한 이용자의 배경지식과 적합성 판정의 관계가 두드러졌다. 또한, 적합문서의 숫자가 적었을 때 이용자들은 적합성 판정에 더욱 어려움을 겪었다. 마지막으로, 검색결과리스트중 상위 N 문서의 적합성 판정에 대한 중요성을 확인할 수 있었다.

Keywords : precision, relevance judgement, information retrieval system evaluation, user behavior.
정확률, 적합성 판정, 정보검색시스템 평가, 이용자행태

* Texas Woman's University (hlee@mail.twu.edu, hyukjinl@rci.rutgers.edu)

■ Received : 15 May 2006
■ Accepted : 22 June 2006

1. Introduction

In the field of experimental information retrieval, performance of information retrieval systems is often measured by precision, which is the percentage of retrieved documents which are actually relevant to the query for which they were retrieved. This measure is controversial, in that relatively small changes in precision in experimental settings (e.g. from 28% to 32%) are often interpreted as "significant", yet we do not know whether such a change would be perceived as significant by a human user of an information retrieval system. The general problem here is that it is not known, what degree of difference in precision is actually perceptible and significant for users of interactive information retrieval systems. This project attempts to determine what level of difference in precision actually makes a difference to human beings. The overall goal is to use these results to inform the evaluation of information retrieval (IR) systems in general.

2. Literature Review

Considering the importance in the point of the evaluation of IR systems, it is surprising that the community of information science have given little attention to the study of perceptible level of precision rate.

In Veerasamy and Heikes (1997), their visualization tool which displays document surrogate information enabling set-at-a-time perusal of documents helps users in identifying document relevance quicker by about 20%. It is believed that users consult visualization before they consult the title, thereby not looking at the titles of those apparent non-relevance documents. Results of an experiment evaluating the tool shows that when users have the tool they are able to identify relevant documents in a shorter period of time than without the tool, and with increased accuracy. Interestingly enough, magnitude of time-decrease due to visualization is much higher in the low precision condition than in the high precision condition. It supports Saracevic's argument (1969) that while minimal information is needed to say that a document is non-relevant, much more information is needed to say a document is relevant. The experiment also shows that users with the visualization tool did better in accurate identification of document relevance. The absolute relevance was judged by Textual REtreival Conference (TREC) assessor. However, unlike results of the experiment in time, there was no significant interaction between precision and visualization on accuracy from both high and low precision conditions. Thus the authors concluded that visualization seems to help increase accuracy to the same extent irrespective of the density of relevant documents. However, what if the length of time is fixed for the same experiment in both high and low precision conditions? Because this study applies the same amount of time to all different

precision conditions, there might be interesting interaction between accuracy and precision rate.

Sormunen (2002) introduces a four-point relevance scale and reports the findings of a project in which TREC-7 and TREC-8 document pools on 38 topics were reassessed. The author suggests that the constituency of assessments is difficult to achieve if the assessors feel topic descriptions ambiguous. This is also associated with the confidence of decisions. In the personal interviews, TREC assessors complained that for many topics the title and description presented contradicting relevance criteria or otherwise did not give a solid basis for relevance assessments. New criteria by the assessors were not always in line with the intentions of original assessors. Because solid relevance judgment on each document and clear description for a subject are crucial in terms of setting of our study, the process of selecting the topics and making the instruments was seriously treated. This process is explained in the research design section.

3. Research Question and Hypothesis

The main research point is to discover what difference in precision level is actually perceptible and understandable by people at various baseline-starting points. That is, the degree of difference in precision could be a critical factor influencing on the people's preference of search results from interactive information retrieval systems. It may cast doubt on tendency taking small change in precision as significant one.

Therefore, the following simplified research question is raised ; what level of difference in precision actually makes a difference to human beings? Two hypotheses have been developed based on the research questions. Two main hypotheses based on these measurements are as follows :

H1. There is a significant level of difference in precision, which users can recognize.
H2. There is a user's preference on one side between the left and the right.

3.1 Research Design

3.1.1 Project design

The general design of the project is to recruit participants who are asked whether one of two lists of retrieval results is preferable to the other, with respect to a given search topic. Participants do not the searching themselves, but are shown two lists which have been experimentally constructed, for each of ten search topics. They are interviewed to determine whether they have a preference, and why.

3.1.2 Experimental design

We recruited ten subjects, who have at least more than 4 years experience in information searching and often do conduct

〈Table 1〉 Ten topics

ID	Code	Title	Description
1	212	Intellectual Property Law Violations	What countries have been accused of failing to adequately protect U.S. copyrights, patents and trademarks?
2	217	Extra-terrestrial Life	Reports on the possibility of and search for extra- terrestrial life/intelligence.
3	221	Stopping Drug/Gang Warfare	Steps taken by church, governments, community, civic organizations to halt carnage among youths engaged in drug or gang warfare.
4	227	Accidental Military Deaths	Identify instances and reasons of deaths in the U.S. military caused by other than enemy (e. g., friendly fire, training accidents).
5	235	Legalizing Drugs	What support is there in the U.S. for legalizing drugs?
6	306	African Civilian Deaths	Reports of civilian non-combatants who have been killed in the various civil wars in Africa.
7	324	Argentine/British Relations	Reports of international relations/contacts between Britain and Argentina.
8	326	Ferry Sinkings	Reports of ferry sinkings where 100 or more people lost their lives.
9	331	World Bank Criticism	What criticisms have been made of World Bank policies, activities or personnel?
10	350	Health and Computer Terminals	Is it hazardous to the health of individuals to work with computer terminals on a daily basis?

a search on World Wide Web everyday. Each of them were asked to judge two lists of 30 retrieved document summaries with respect to each of ten topics 〈see Table 1〉. For our data we used a collection of documents which have been retrieved with respect to a large number of search topics, and which have been judged with respect to their relevance to the topics by three different judges. These materials are available to us by virtue of our participation in the NIST-sponsored Text REtrieval Conferences (TREC). Based on such data, we did double-check and judged document relevance to generate the final set of documents for the experiment.

For each topic, there are five list conditions :

1. 30% precision vs. 50% precision ;
2. 40% precision vs. 50% precision ;

〈Table 2〉 Experimental design

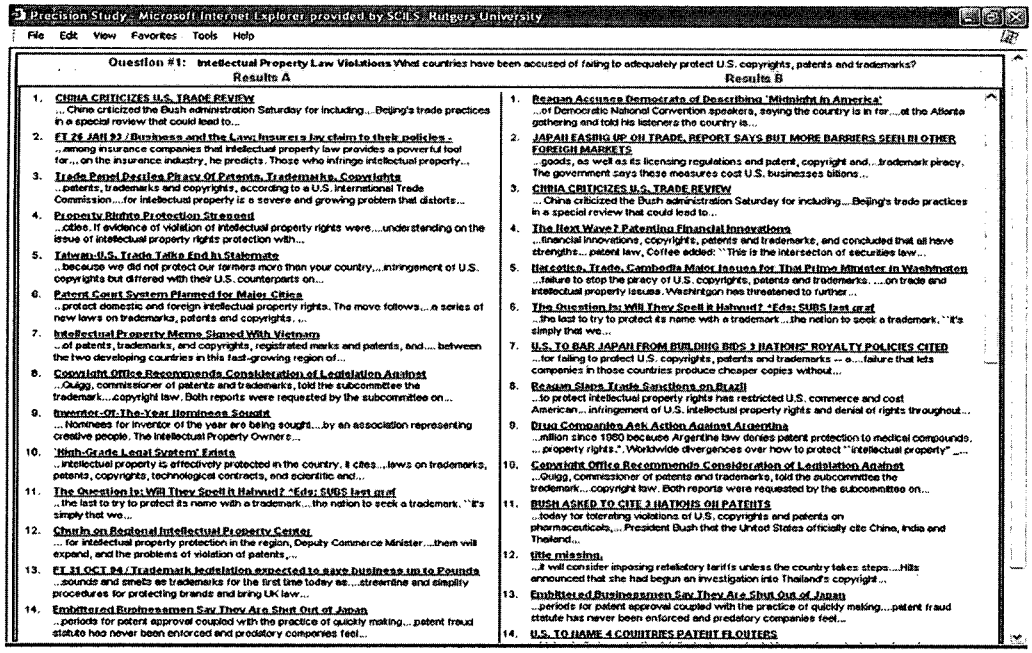
Sub.	Topic-Condition, & Control condition side									
	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right
1	1-1	2-2	3-3	4-4	5-5	6-1	7-2	8-3	9-4	10-5
2	1-2	2-3	3-4	4-5	5-1	6-2	7-3	8-4	9-5	10-1
3	1-3	2-4	3-5	4-1	5-2	6-3	7-4	8-5	9-1	10-2
4	1-4	2-5	3-1	4-2	5-3	6-4	7-5	8-1	9-2	10-3
5	1-5	2-1	3-2	4-3	5-4	6-5	7-1	8-2	9-3	10-4
6	10-5	9-4	8-3	7-2	6-1	5-5	4-4	3-3	2-2	1-1
7	10-1	9-5	8-4	7-3	6-2	5-1	4-5	3-4	2-3	1-2
8	10-2	9-1	8-5	7-4	6-3	5-2	4-1	3-5	2-4	1-3
9	10-3	9-2	8-1	7-5	6-4	5-3	4-2	3-1	2-5	1-4
10	10-4	9-3	8-2	7-1	6-5	5-4	4-3	3-2	2-1	1-5

3. 50% precision vs. 50% precision ;
4. 60% precision vs. 50% precision ;
5. 70% precision vs. 50% precision.

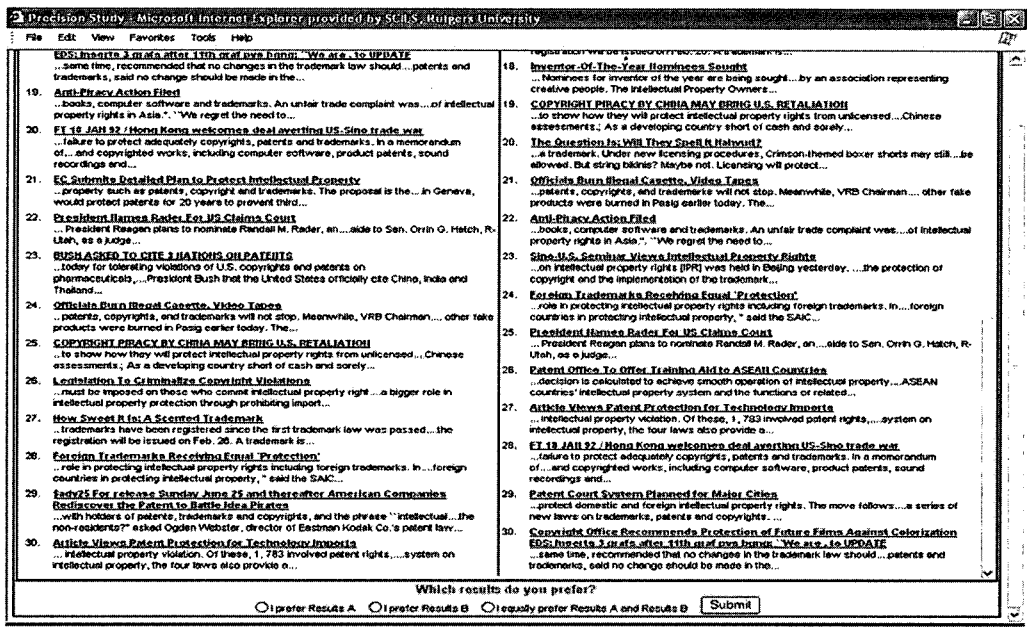
We choose 50% as the baseline precision, as this is currently representative of the best-performing information retrieval systems. Precision are varied as above. Each list begins with a relevant document, and then is systematically "seeded" with relevant documents according to the given precision level (interspersed with documents which have been judged to be not relevant). In other words, the list is not ordered by any ranking technique. All participants judged ten topic-condition pairs, with two instances of each condition. Each topic-condition pair was evaluated by two different participants, and the order of the pairs for the participants was rotated according to condition. There are two topic orders, one from 1 to 10, the other from 10 to 1. For each topic-condition pair, the control (50%) condition appears on the right once and on the left once. The final design is indicated in 〈Table 2〉.

3.1.3 Interface

System interface is simplified and clear for the goal of the study. A given topic is located at the top of the interface. Two parallel document lists lie in the body of the screen. Each document in the list was represented by a title and the three "most descriptive" sentences of that document, with respect to the topic. Any time a subject may open the document by clicking a specific document title. A scroll bar enables a subject to reach the bottom of the two lists and he/she makes a decision among three clickable buttons : List A (left), List B (right), and No preference.



〈Figure 1〉 System Interface (top)



〈Figure 2〉 System interface (down)

Once a subject completes his/her decision, the next topic is given with new lists 〈see Figure 1 and 2〉.

3.1.4 Experimental procedure

On arrival at the study site, participants were administered the informed consent form. Then they completed a questionnaire in which they were asked to evaluate their fluency in written and spoken English, and their previous experience and familiarity with information retrieval systems. They were then be given detailed instructions concerning what they would be doing during the course of the study. Then they were given a sheet of paper which has a description of the first topic, and completed a brief questionnaire on their evaluation of their knowledge of the topic, and their confidence in their evaluation. The subjects then moved to a monitor on which were displayed the two lists which they were to compare, in two scrollable windows. Think aloud method was not used while examining the lists to minimize distraction or inhibition. The subjects examined these lists for up to ten minutes, and then were asked to state whether they have any preference between the two lists, and if so, which one they prefer. They were also asked to explain why they made this decision, and what difficulties they had in making it. They were then be given the second topic, and the same procedure was followed. This was repeated for all ten topics. After the tenth topic procedure was complete, they were interviewed concerning their familiarity with and response to the method of presentation of the results. The screen capturing software was used, while there was no video or audio taping.

3.2 Findings

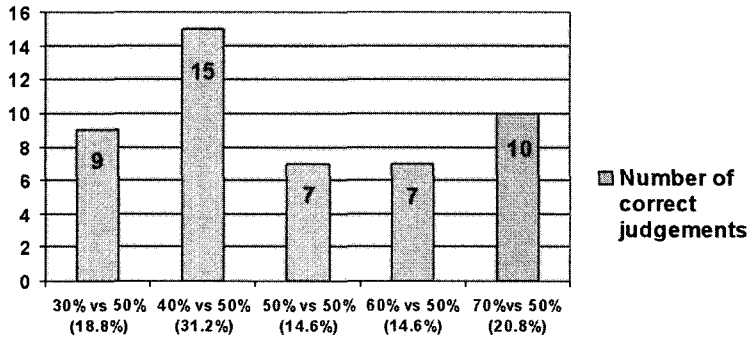
3.2.1 Significant level of difference in precision

Condition 1. 30% precision vs. 50% precision ; 9 (18.8%)
Condition 2. 40% precision vs. 50% precision ; 15 (31.2%)
Condition 3. 50% precision vs. 50% precision ; 7 (14.6%)
Condition 4. 60% precision vs. 50% precision ; 7 (14.6%)
Condition 5. 70% precision vs. 50% precision 10 (20.8%)

Right answer is 48% and the range of the number of right answers among the subjects is from 3 to 7in total 10. The range of the number of right answers among the topics is also from 2 to 8 in total 10. The lowest correct answer is 14.6% for the condition 3 and 4 and the highest correct answer is 31.2% for the condition 2 〈see Table 3 and Figure 3〉. Condition 1 and 5 are relatively high ; 18.8% and 20.8%. A one-sample chi-square test was conducted to assess whether a subject could judge right in a condition having a specific difference level of precision. The results of the test were not significant, $2(4, \underline{N} = 48) = 4.50$, $\underline{p} = .343$. Although a follow-up test also indicated that the proportion of subjects who were correct in condition 2 did not

〈Table 3〉 Cumulated judgments for best list

Condition 1	Condition 2	Preferred 1	Preferred 2	No preference
30% VS. 50%		5	9	6
40% VS. 50%		1	15	4
50% VS. 50%		4	9	7
60% VS. 50%		7	7	6
70% VS. 50%		10	3	7



〈Figure 3〉 Judgment among different conditions

differ significantly from the proportion of subjects who were correct in condition 3 or 4 which the least of participants were correct, $2(1, N = 22) = 2.91$, $p = .088$, the result was moderately meaningful.

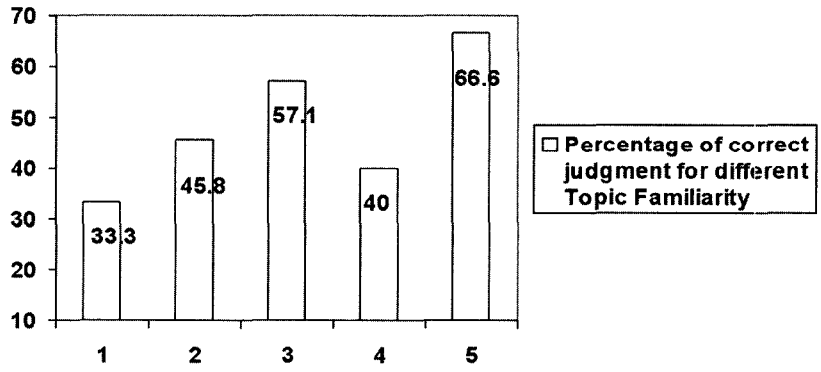
In average, 5 minutes and 8 seconds has been taken for each topic judgment. There was no relationship between time and correctness of relevance judgment (see Appendix, p. 18). Six English native speakers and four non-native speakers participated for the study. While, the average taken time for each task for English natives (N=6) is 4 minutes and 44 seconds, the one for non-natives (N=4) is 5 minutes and 46 seconds. However, the average numbers of correct judgments are not much different between two groups ; 4.83 for English native group, and 4.75 for non-native group.

〈Table 4〉 Cumulated judgments for side

Correctness	Left	Right	No preference
Correct	22	19	7
Incorrect	12	17	23
Totals	34	36	30

〈Table 5〉 Topic familiarity and judgments for best list

Familiarity score	Number of familiarity scores	Number of correct judgment	Average rate of correct judgment
1	21	7	33.3%
2	24	11	45.8%
3	42	24	57.1%
4	10	4	40%
5	3	2	66.6%
Total	100	48	100%



〈Figure 4〉 Percentage of correct judgment for different Topic Familiarity

3.2.2 User's preference on selecting aside between the left and the right

Basically, there is no apparent preference to one side over the other between the left and the right list 〈see Table 4〉. Among 100 cases, 34 for the left side, 36 for the right side, and 30 are for no preference. There is no preference for either side when a subject was correct or incorrect for relevance judgment.

3.2.3 User's familiarity to topic and correctness in precision judgement

Interesting result is found in the relationship between the subjects' topic familiarity (pre-knowledge level) and rate of correct judgments 〈see Table 5〉. There is a quiet clear correlated relation between the two variables; more familiar to the topic, more correct relevance judgment 〈see Figure 4〉. If we exclude topic familiarity level 4 and 5 which comprise only 13% of the total cases (N = 100), the

relationship is more apparent.

3.2.4 Agreement on relevance judgment

As explained in research design section, each topic-condition pair was evaluated by two different participants. The agreement between two judges is investigated. As indicated in 〈Table 6 and 7〉, the rate of agreement between two judges is low by both topic and condition. Overall, only 24% agreement occurred and even when they did agree, just 46% (11/24) were correct in relevance judgment task. However, in condition2, all subjects who agreed on their judgment were correct (5/5).

〈Table 6〉 Agreement on judgments, by topic

Topic	Agreement	Number Correct
1	2	0
2	3	0
3	3	2
4	2	1
5	2	1
6	3	0
7	1	0
8	3	3
9	2	2
10	3	2

〈Table 7〉 Agreement on judgments, by condition

Condition	Agreement	Number Correct
1	5	2
2	5	5
3	6	2
4	5	1
5	2	1

3.3 Questionnaire analyses

3.3.1 Regarding relevance judgment

Subjects were asked to explain how they were confident of their decision and why they made their decision, and what difficulties they had in making it.

First, most subjects were quite confident on their decisions despite just 48 % of total correctness. A few of them showed strong confidence on their decision in the condition 2 (40% vs. 50%), which is interesting reflecting the result that in this condition people decided correct judgments best. Often, the subjects were confident equally for not only their right judgments but also wrong ones.

Another finding is the importance of the first top N documents in a list for relevance judgment task. More than half of ten subjects confessed that although they knew that the list is not rank-ordered from the instruction, the location of relevant documents influenced their decision seriously. There was no agreed number for the top documents among the subjects but no more than 10 ; 3, 4, 5, and 10 document number were mentioned as the top document numbers which mattered in relevance judgment. This tendency was even clearer from the result that even in

the condition 3 (50% vs. 50%), one subject preferred one list based on first 4 documents in the list 2 of 4 relevant documents lied in one list, none in the other. Another subject also decided the left list even though she already recognized that both lists were almost same ; because the relevant documents were located higher in the left list than the right one. She even indicated that the right list had very relevant one at the bottom.

Finally, it turned out that the subjects feel more difficult in a situation they judge between the two lists having more non-relevant documents than in a situation they do the lists having more relevant documents. In addition, when there are more overlapped (duplicated) documents between the two lists, majority of the subjects have trouble in judging. Some subjects felt difficulty of lack of knowledge of a topic as indicated in the finding section.

3.3.2 Regarding document representation

From the exit interview, we tried to find how people feel the way of document presentation in a list ; each document in the list was represented by a title and the three most descriptive sentences of that document with respect to the topic.

Most subjects depended on three sentences of display for their judgments. One subject even confessed, "I did not want to open full-text as much as I could". Another subject said that it depends on the topic task but this method gives enough general idea on judgment. Only one subject gave the least scores for 5 scale Likert measurement about this way of displaying documents, but most of the subjects showed that they were felt very familiar (4.56 of 5) and felt useful (3.56 of 5) to the way of representing a document in this study. All subjects explained that they felt comfortable about the display because it is similar to what they can see in the Web search engine display the most frequent mentioned was the Google. Some suggestions for additional information for the display of the document are : date, resource information, and highlighted keywords.

There are some interesting points but these were mentioned by single subjects. One subject mentioned that he was affected by reading order from reading the left list to the right because he has gained some topic familiarity through the process skimming the first list. Another subject suggested that if it is about 'aspect' type of lists, not document type, it would be more ideal in terms of comparison type task.

4. Discussion and Future Study

Even though it was a disappointing result that people could not recognize the different precision rates significantly ; it was interesting that they were very prominent in one specific precision condition, 40% vs. 50%. The result was

much higher than the two conditions having 20% precision difference and the other condition with 10% difference, 60% vs. 50%. Here we cannot define an apparent reason for this but it seemed that people were even confident on their decisions. We think this phenomenon is worth to be investigated more seriously in further studies. Furthermore, despite statistically non-significant, two conditions having 20% precision difference (30%, 70 % vs. 50%) are moderately higher in precision correctness than the condition having no precision difference between both lists (50% vs. 50%).

It is valuable result that the correctness of relevance task has little to do with the taken time for the task. It is clearer from the result comparing the two groups of subjects, English native and non-native. Despite clear different average time between the two groups, the average numbers of correct judgments are not much different. We also investigated if there is any apparent preference to one side over the other between the left and the right list. Interestingly there is no preference for either side when a subject was either correct or incorrect for relevance judgment. In conclusion, the location of the information retrieval result does not matter for users at least in their cognitive preferences.

The strong relationship between the subjects' topic familiarity (pre-knowledge level) and rate of correct judgments is one of the most interesting results in this study. While neither the taken time nor location of the document list did affect user's correct relevance judgment, the topic familiarity did support their judgments much. This fact gives very potential idea on future study. If we can control the topic familiarity level so that all subjects have no difference in terms of topic familiarity for their tasks, they might have a different tendency toward different precision conditions.

The random result of the agreement between two judges for each topic-condition pair by two different participant suggests that in future studies regarding this issue, it may be necessary to have more controlled subject group (e. g. the group with the similar level of the topic tasks). However, once more, very interestingly the subjects who agreed on their judgments in the precision condition, 40% vs. 50%, were all correct.

It turned out that the subjects have more difficulty in a situation they have to judge between the two lists having more non-relevant documents than in a situation they do between the lists having more relevant documents. It could be related to several issues such as the level of information task difficulty or the method of representing a document. Even though the units of analyses in previous studies were diverse ; for example, title, abstract, and full-text (Saracevic 1969), the basic arguments explained that less information is needed to say that a document is non-relevant, much more information is required to say a document is relevant. Therefore, such opposite result from our study is

interesting and seems to be worth to be investigated further.

We could confirm that information users, especially at least the ones who has relatively enough experience, are very subjective with high self-esteem on their decision on relevance task. From the result and interview, most of subjects showed quite confidence on their decisions regardless of the correctness.

Finally, in displaying document point of view, the serious influence from the first top N documents in a list for relevance judgment task has been confirmed ; more than half of subjects confessed that the location of relevant documents influenced their decision seriously with no regards to the instruction indicating the non-rank ordered list. High precision rate in top N documents is one of most crucial factors for information users to have preference and confidence on the list. In addition, most subjects felt comfortable and useful three sentences of display for their relevance judgments mainly because it is similar to what they have used in the Web search engine display. A user may want to minimize cumbersomeness of opening full-text documents by the efficient display of the document contents. A few suggestions for additional information for the display of the document are the information on date, resource information, and highlighted keywords however, nobody seriously agreed the necessity of metadata menu bar.

5. Conclusion and Limitation

Considering its importance for the evaluation of information retrieval (IR) systems in general, there have been little attempts to reveal the perceptible level in precision rate and further evaluation from cognitive angle of a ranking list. In this point, we regard this study is valuable and some interesting findings should be useful for other related information retrieval researches. The limitation of the study is the small sample size ; only ten subjects participated in this study and their levels of English were different. Future study should have more subject number (and more topic number) so that it can have more statistical validity.

References

ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97), 236-245.

NIST Special Publication 500-236. 1995. The Fourth Text REtrieval Conference (TREC-4) : *the proceedings of the fourth Text REtrieval Conference (TREC-4)*

(Gaithersburg, Maryland, November 1-3.)

NIST Special Publication 500-240. 1997. The Sixth Text REtrieval Conference (TREC-6) : *the proceedings of the sixth Text REtrieval Conference (TREC-6). (Gaithersburg, Maryland, November 19-21.)*

Saracevic, T. 1969. "Comparative effect of titles, abstracts and full texts on relevance judgments." *Proceedings of the American Society for Information Science, 6*, 293-299.

Sormunen, E. 2002. "Liberal Relevance Criteria of TREC - Counting on Negligible Documents? In Beaulieu, M. et al. (Eds.)" *Proceedings of the Twenty-Fifth Annual ACM SIGIR Conference on Research and Development in Information Retrieval (Tampere, Finland August 11-15.), Special Issue of SIGIR Forum, 36*, 324-330.

Veerasamy, A. & Heikes, R. 1997. Effectiveness of a graphical display of retrieval results. In Belkin, N.J. et al., (Eds.) : *Proceedings of the 20th Annual International*

Appendix. Correctness of judgment and Time

	T1 (L)	T2 (R)	T3 (L)	T4 (R)	T5 (L)	T6 (R)	T7 (L)	T8 (R)	T9 (L)	T10 (R)	Total & Time
Sub1	W 8:16 (N,R)	R 7:36 (L,L)	R 7:37 (N,N)	R 6:03 (R,R)	R 7:32 (L,L)	R 4:06 (L,L)	R 6:38 (R,R)	W 4:28 (L,N)	W 6:47 (N,L)	W 6:41 (N,R)	6 6:34
Sub2	W 5:36 (L,R)	W 5:55 (L,N)	R 3:55 (L,L)	R 5:41 (R,R)	R 5:00 (R,R)	R 3:40 (L,L)	W 5:45 (R,N)	R 5:59 (R,R)	R 6:57 (L,L)	R 4:52 (L,L)	7 5:26
Sub3	W 9:16 (R,N)	W 6:59 (N,R)	R 6:44 (L,L)	W 6:57 (N,L)	W 5:40 (N,R)	W 9:52 (R,N)	W 11:26 (R,L)	R 8:19 (R,R)	R 8:28 (R,R)	R 7:08 (L,L)	4 8:05
Sub4	R 7:08 (L,L)	W 7:05 (L,R)	W 6:28 (N,R)	R 5:42 (L,L)	R 5:17 (N,N)	W 4:15 (N,R)	R 6:16 (L,L)	R 6:14 (L,L)	R 6:59 (R,R)	W 10:32 (R,N)	6 6:40
Sub5	W 5:36 (N,L)	W 3:04 (R,L)	R 3:54 (R,R)	W 4:40 (R,N)	W 3:15 (R,L)	W 5:45 (L,R)	R 2:31 (R,R)	R 3:28 (L,L)	R 3:20 (N,N)	W 5:02 (L,R)	4 4:04
	T1 (R)	T2 (L)	T3 (R)	T4 (L)	T5 (R)	T6 (L)	T7 (R)	T8 (L)	T9 (R)	T10 (L)	
Sub6	W 5:25 (N,L)	W 4:26 (N,R)	W 2:07 (R,N)	W 4:56 (N,L)	W 4:08 (N,R)	W 4:06 (N,R)	W 4:14 (N,L)	R 2:51 (N,N)	R 3:35 (R,R)	R 4:52 (L,L)	3 4:04
Sub7	R 2:51 (L,L)	W 6:14 (L,N)	R 3:31 (R,R)	W 4:52 (N,L)	W 6:17 (R,L)	W 2:55 (N,R)	R 2:52 (N,N)	W 5:52 (R,L)	W 5:57 (N,R)	R 1:43 (R,R)	4 4:18
Sub8	W 6:24 (L,N)	W 9:22 (N,L)	W 5:18 (N,R)	R 12:23 (R,R)	R 2:28 (L,L)	W 2:10 (R,N)	W 3:40 (N,R)	R 3:32 (L,L)	W 3:37 (R,L)	R 4:57 (R,R)	4 5:23
Sub9	W 2:35 (L,R)	R 3:21 (L,L)	W 1:33 (R,L)	R 3:07 (R,R)	R 2:49 (N,N)	W 6:58 (R,L)	W 0:20 (L,R)	R 0:43 (R,R)	R 2:46 (L,L)	W 2:10 (R,N)	5 2:38
Sub10	R 5:35 (R,R)	W 5:56 (L,R)	R 2:08 (L,L)	W 2:21 (R,N)	W 6:59 (L,R)	W 3:28 (N,L)	W 3:26 (N,L)	R 4:39 (R,R)	R 3:47 (N,N)	R 4:14 (L,L)	5 4:15
Total	3 5:52	2 6:00	6 4:20	5 5:40	5 4:57	2 4:44	4 4:43	8 4:37	7 5:13	6 5:13	48 5:08

R: Right answer
W: Wrong answer
(A: Subject's selection, B: Correct information)
Time: Minute and Second