Directed Novelty and Redundancy in Information Retrieval

Joseph Tan Kai Huang[1] and Xu Yunjie Calvin[2]

School of Computing, National University of Singapore
3 Science Drive 2, Singapore 117543

ABSTRACT

The information science research community is divided between the system-centered cluster working on information retrieval (IR) algorithms and the user-centered cluster working on user behavior. A criticism from the user-centered cluster on the system-centered cluster is that document relevance should be regarded as a subjective, dynamic, and multi-dimensional construct, rather than a static and objective one. Building on past user studies, this study proposes two hypotheses: firstly, the uni-dimensional treatment of relevance in current relevance feedback IR systems will perform as well as a feedback system based solely on topicality judgment. Secondly, an IR system, which takes into account users' novelty perception, will perform better[3] when the novelty judgment is allowed to age – some documents are regarded as redundant if they have been evaluated long ago – than when redundancy is not incorporated into novelty judgment. Four systems were designed to test these hypotheses and results were collected from the simulation and user study tests. The results of the simulation support the two hypotheses whereas the results of the user study support only the first hypothesis and not the second, which may be due to the experimental design.

1 INTRODUCTION

There is a split in paradigm in the research community for information science: on one side is the system-centered approach on IR algorithms while on the other side, a user-centered approach on user's behavior. Bridging the gap between the user-centered and system-centered approach towards IR algorithms has indeed remained a challenge to the information science research community.

On the user-centered side, a significant amount of research has been done on identifying the factors that influence users' perception of the relevance of a document. In this study, relevance is defined as the match between the user's information need and the document. An information need is the general topic of interest in a search session. Out of these factors, Xu and Chen [1] sieved out two major factors: topicality being the most important factor followed by novelty. Topicality is the extent to which a document is about the topic of interest and novelty is the extent to which the content of a retrieved document is new to the user or different from what the user has known before. Separate from user-centered studies, system-centered studies have also suggested considering novelty and topicality during relevance judgment of a document. In this study, relevance judgment of a document is defined as the user's perceived degree of relevance. One such a study is the research done by Zhang [2]. In his research, Zhang included in relevance judgment the user's judgment of both topicality and novelty of a document, the former being called topicality judgment and the latter novelty judgment. He claimed that the novelty of a document should be defined as the amount of relevant information in the current document that is not covered by relevant read documents. Such novelty is termed as undirected novelty. Undirected novelty is also known as the opposite of redundancy. Hence, redundancy is the amount of relevant information in the current document that is already covered by relevant read documents. Yin [3], however, criticized Zhang's assumption of

[1]Student
[2]Supervisor
[3]An IR system's performance depends on its performance in relevance judgment of an unseen document

users completely assimilating the information in a document which is retrieved by an IR system that uses undirected novelty in relevance judgment of an unseen document. He argued that users might want to find out more about a sub-topic to fulfill their information needs. Therefore, his study revealed that an IR system should use directed novelty, which is the amount of relevant information that contains novel aspects of a topic, and be integrated together with topicality in a stepwise fashion, which is to first eliminate the documents by topicality, then sort the remaining documents by novelty.

This study seeks to further augment IR algorithms with users' novelty judgment. Noting that relevance judgment is biased towards topicality, it has been hypothesized in Yin's paper that relevance judgment in a Rocchio feedback system[4] will give a similar result in a Rocchio feedback system based on topicality judgment. This paper sets out to test this hypothesis. Furthermore, his research revealed that when undirected novelty is combined with topicality, undirected novelty causes worse relevance judgment of an unseen document as compared to directed novelty. Since undirected novelty is the opposite of redundancy, it seems that there is no need in relevance judgment to include redundancy as including undirected novelty deteriorates relevance judgment. However, redundancy can be used in relevance judgment as users may shift to another sub-topic during their search and we can use the concept of redundancy to regard documents of past sub-topics as redundant. Therefore, in my paper, I will test the second hypothesis that there will be better relevance judgment by incorporating redundancy into directed novelty during relevance judgment.

2 IR SYSTEMS

Four different systems are designed to test out the hypotheses. The four systems are Rocchio Feedback System – Relevance (RFR), Rocchio Feedback System – Topicality (RFT), Directed Novelty-Topicality Feedback System (NT) and Direct Novelty-Topicality-Redundancy Feedback System (NTR). Each system will use user's perception of past documents to make relevance judgment of an unseen document. The systems will be tested against each other to test out the hypotheses. RFR and RFT are used to test the first hypothesis. As RFT, a topicality feedback-based system (which is called topicality feedback system thereafter), only relies on the topicality profile and RFR, a relevance feedback-based system (which is called topicality feedback system thereafter), only relies on the relevance profile, if RFR and RFT performs equally well, it will support the first hypothesis. A relevance profile is a vector of terms which represents the user's dynamic information need. A topicality profile is a vector of terms which represents the user's topicality perception whereas a novelty profile is a vector of terms which represents the user's novelty perception and the novelty can be considered directed or undirected. NT and NTR will be used to test the second hypothesis. NT is chosen as a benchmark for NTR as NT is the best performing IR algorithm. NT and NTR will be used to test out the second hypothesis as NTR ages the directed novelty profile while NT does not. If NTR performs better than NR, the second hypothesis is supported.

2.1 Rocchio Feedback System – Relevance (RFR)

RFR is a relevance feedback system based on the Rocchio system and uses only the relevance profile to make relevance judgment of an unseen document. It will be used to test against RFT to prove the first hypothesis. The relevance profile is used to capture the user's relevance perception. A user's relevance profile starts with his initial query and is subsequently built according to the user's

[4] A Rocchio system is one that improves the query using the user's evaluation of the relevance of the document.

evaluation of the relevance of documents. During the construction of the profile, TFIDF[5] weighting is used because the profile uses a vector of terms which is largely biased towards topicality. Also, the relevance profile is used to make relevance judgment of an unseen document in RFT. The relevance profile updating strategy is based on Rocchio's relevance feedback. If a user evaluates and assigns relevance scores to a set of documents, the relevance scores can be used to update the initial profile and thus retrieve subsequent unseen documents. Relevance judgment of an unseen document is calculated using the cosine similarity between the relevance profile and the document vector.

2.2 Rocchio Feedback System – Topicality (RFT)

The second system, RFT, is a variant of the Rocchio system. It is a topicality feedback system, instead of a relevance feedback system. RFT uses only the topicality profile to make relevance judgment of an unseen document. The topicality profile differs with the relevance profile by using topicality scores instead of relevance scores for its updating strategy and judgment.

2.3 Directed Novelty-Topicality Feedback System (NT)

The next system, NT, is an IR system which integrates the directed novelty profile and topicality profile in a stepwise fashion during relevance judgment of an unseen document. To measure the directed novelty profile, we will use the probabilistic measure F4. The F4 measure of a term is the ratio of relevance odds and non-relevance odds, which is the ratio of the odds that a relevant document contains term t_j and the odds that an irrelevant document contains t_j. This study will use a variant of the F4 measure by using the F4 measure to identify novelty-feature terms – terms that best differentiates novel documents from the non-novel ones – in a set of novel documents and a set of non-novel documents. With directed novelty profile, a document's directed novelty can be calculated with the cosine score between the document vector and the directed novelty profile. As Yin [3] suggested that the directed novelty profile should be integrated with the topicality profile in a stepwise fashion, this system will do so.

2.4 Directed Novelty-Topicality-Redundancy Feedback System (NTR)

The last system, NTR, is an IR system which integrates the directed novelty profile and topicality profile in a stepwise fashion during relevance judgment of an unseen document. On top of that, it takes into account the ageing of the directed novelty profile. The redundancy profile can remove redundant sub-topic document which does not follow current sub-topic documents. The redundancy profile is a vector of terms to represent the user's redundancy perception. The topicality and directed novelty profile in NTR follows those used in NT. The redundancy profile has been used in calculating the undirected novelty profile as undirected novelty is defined as the opposite of redundancy. To build the redundancy profile, the maximum marginal relevance model is used, whereby the marginal relevance of a document is measured with a weighted sum of its similarity to the query and its redundancy to previously selected documents. We apply the idea to our context. The redundancy profile will first be used to age the directed novelty profile by subtracting the novelty score of each document by the redundancy score, and the combined profile will then be integrated

[5] TFIDF, or term frequency-inverse document frequency, is a popular scheme for weighting terms. It divides each term frequency weight with its frequency of the term in the same language. It is to reduce the importance of the term if it is common in the same language, such as stop words.

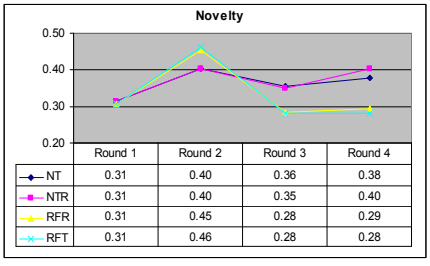
with the topicality profile in a stepwise fashion. In this study, we will start ageing the directed novelty after the second round of evaluation of the documents.

3 EXPERIMENTAL SET-UP

The experiment will use the four systems explained above to test out the two hypotheses. Understanding the experimental set-up for the user study is essential not only for the user study, but also the simulation as the user data used in the simulation was produced by an experiment of similar set-up done by Yin [3]. The experiment design for this user study will follow closely to the one set up by Yin [3]. The search topic for this study was "mobile phone radiation and health". A corpus for the experiment was collected from Google's search engine. In the end, 295 documents were used for the experiment. The subjects were randomly assigned to an IR system. However, all systems had an identical user interface. The users were given a web address to access the system. After accessing the system, they were brought to a page with a text box and a search button. The initial query in the text box was "mobile phone health". When the query is submitted, the 10 most relevant documents were returned and listed in one page with only the document title and evaluation boxes for the topicality, novelty and relevance. All the subjects had to click on the title, which would link them to the document. They had to read the document and evaluate the three attributes on an eight-point scale for every document. The subject would then click the "next" button to move to the next page. The system would then retrieve another 10 most relevant unread documents according to the subjects' evaluations and the system they were assigned to. The process continued until the subjects completed six rounds.

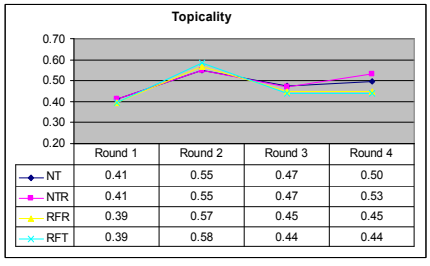
4 SIMULATION

Before the user study was done, a simulation was done with the four IR systems. The simulation served as a preliminary test towards the two hypotheses and to see the viability of conducting a full-scale user study. NR and RFT has been tested in Yin's [3] user study. Thus, I used the topicality, novelty and relevance scores of evaluated documents collected from Yin's user study. For the simulation, NTR used NR's scores of evaluated documents and RFR used RFT's scores of evaluated documents. The systems were simulated with four rounds, instead of six rounds used in the user study. The simulation was done by feeding in the user's document scores into the system. The three diagrams below shows the average scores among all users simulated on for each round for novelty, topicality and relevance. It shows the variation of scores between the different rounds. Figure 1 reports the results for novelty whereas Figure 2 reports the results for topicality and Figure 3 for relevance. I normalized the score into a range of 0 to 1 by dividing scores with the maximum score of 7. For each table, the average scores is reported for each round and system.



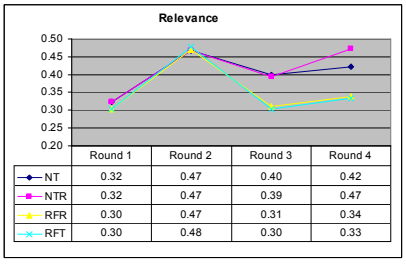
	Round 1	Round 2	Round 3	Round 4
NT	0.31	0.40	0.36	0.38
NTR	0.31	0.40	0.35	0.40
RFR	0.31	0.45	0.28	0.29
RFT	0.31	0.46	0.28	0.28

Figure 1. Simulation Results for Novelty



	Round 1	Round 2	Round 3	Round 4
NT	0.41	0.55	0.47	0.50
NTR	0.41	0.55	0.47	0.53
RFR	0.39	0.57	0.45	0.45
RFT	0.39	0.58	0.44	0.44

Figure 2. Simulation Results for Topicality



	Round 1	Round 2	Round 3	Round 4
NT	0.32	0.47	0.40	0.42
NTR	0.32	0.47	0.39	0.47
RFR	0.30	0.47	0.31	0.34
RFT	0.30	0.48	0.30	0.33

Figure 3. Simulation Results for Relevance

T-test between RFR and RFT (p)				
	Rounds			
Comparison	*2*	*3*	*4*	*2,3,4*
Novelty	0.18	0.71	0.12	0.53
Topicality	0.019	0.16	0.35	0.94
Relevance	0.34	0.23	0.58	0.67

Figure 4. T-test Results (RFR and RFT)

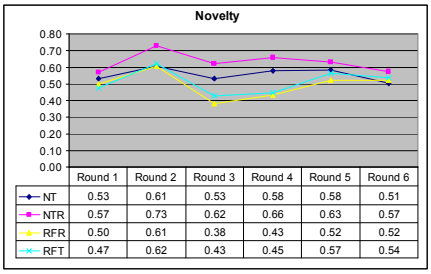
T-test between NT and NTR (p)			
	Rounds		
Comparison	*3*	*4*	*3,4*
Novelty	0.49	0.014	0.095
Topicality	0.59	0.00049	0.011
Relevance	0.43	0.00010	0.0050

Figure 5. T-test Results (NT and NTR)

A two-tailed paired T-test was done on the results and the T-tests results are shown above. For the comparison between RFR and RFT as shown in Figure 4, all the probabilities are above 5%, except for topicality in Round 2. This indicates that there is no significant difference in scores between RFR and RFT – they have similar performance. Thus, the simulation results support the first hypothesis. For the comparison between NT and NTR as shown in Figure 5, the tests indicated that NTR has made a significant difference over NT in novelty, topicality and relevance in Round 4, but NTR has a significant impact only on topicality and relevance when testing both Round 3 and 4 together. Since NTR has a significant improvement on the relevance judgment, as it has an improvement in relevance as seen in Figure 3, over NT, the simulation results supports the second hypothesis. Therefore, the simulation supports both hypotheses and gives us enough grounds to proceed with the user study.

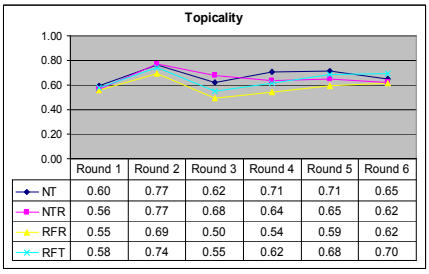
5 USER STUDY

The user study was held in a computer laboratory. The subjects were randomly assigned to the four IR systems. The users were 85 undergraduate students from National University of Singapore. The experiment was carried in multiple sessions, with about 15 students in each session. They were paid SGD$15 after completing the experiment. To motivate the users to take the experiment seriously, an additional SGD$50 was awarded to the user who scored the highest in each session for the post-experiment quiz. The three diagrams below show the average scores among all users for each round for novelty, topicality and relevance. Figure 6 reports the results for novelty whereas Figure 7 reports the results for topicality and Figure 8 for relevance. I normalized the score into a range of 0 to 1 by dividing actual user score with the maximum score of 7. For each table, the average scores is reported for each round and system.



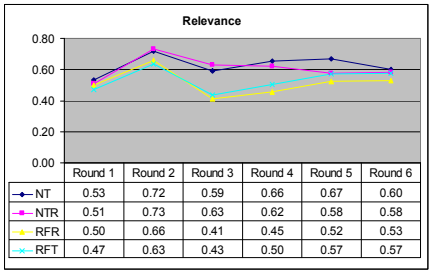
	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
NT	0.53	0.61	0.53	0.58	0.58	0.51
NTR	0.57	0.73	0.62	0.66	0.63	0.57
RFR	0.50	0.61	0.38	0.43	0.52	0.52
RFT	0.47	0.62	0.43	0.45	0.57	0.54

Figure 6. User Study Results for Novelty



	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
NT	0.60	0.77	0.62	0.71	0.71	0.65
NTR	0.56	0.77	0.68	0.64	0.65	0.62
RFR	0.55	0.69	0.50	0.54	0.59	0.62
RFT	0.58	0.74	0.55	0.62	0.68	0.70

Figure 7. User Study Results for Topicality



	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
NT	0.53	0.72	0.59	0.66	0.67	0.60
NTR	0.51	0.73	0.63	0.62	0.58	0.58
RFR	0.50	0.66	0.41	0.45	0.52	0.53
RFT	0.47	0.63	0.43	0.50	0.57	0.57

Figure 8. User Study Results for Relevance

Using the analysis of covariance (ANCOVA), I compared the average novelty, topicality and relevance scores of the four systems. Comparing RFR and RFT, the results suggests that the difference in topicality ($F_{1,211} = 3.02$, $p = 0.084$, $\eta^2 = 0.014$) is significant, but not relevance ($F_{1,211} =$

3.24, $p = 0.073$, $\eta^2 = 0.015$) and novelty ($F_{1,211} = 4.25$, $p = 0.040$, $\eta^2 = 0.020$). Therefore, the results support the first hypothesis as the both systems perform equally well. As for NT and NTR, the differences in relevance ($F_{1,164} = 0.86$, $p = 0.35$, $\eta^2 = 0.0052$), novelty ($F_{1,164} = 0.17$, $p = 0.68$, $\eta^2 = 0.0010$), and topicality ($F_{1,164} = 0.30$, $p = 0.59$, $\eta^2 = 0.0018$) are insignificant. NT is insignificantly better than NTR in relevance and topicality, but not novelty, where NTR is insignificantly better. Thus, the second hypothesis is not supported in this study. This may be due to various reasons. The directed novelty could have been aged too fast, giving users too little time to focus on a certain sub-topic before moving to another sub-topic. Also, there may be too few rounds of documents to show the improvement. On the other hand, it could be due to directed novelty having to age after two rounds of documents and not having catered to the individual user's aging speed. Each user may shift to other sub-topics at different speeds – some in the third round while some in the fourth round. This could have affected the results. Another probable reason would be that the experimental design might inhibit any significant improvements NTR could have given; the search topic and document set might be too narrow for users to do a substantial shift in sub-topics. A possible future research can be done by testing IR systems which directly download documents from the Internet. Broader topics can also be given for the experiment. This expands the user's document set and sub-topics.

6 CONCLUSION

We can see that although relevance is multi-dimensional, it can be reduced to two major factors – topicality and novelty. Current relevance feedback system is not able to capture the multi-dimensionality of relevance – relevance feedback systems are biased towards topicality and this paper supports such an assertion. Furthermore, Yin's study [3] revealed that directed novelty will fare better than undirected in relevance judgment, thus discounting the need for a redundancy profile used in undirected novelty. This paper brings his research further by setting out to test whether the redundancy profile can be used to age the directed novelty profile, thus bringing better relevance judgment. However, results from the user study have shown that such addition did not bring about significant improvement in the relevance judgment. This might be due to the need for a better experiment design to test this algorithm. In conclusion, this paper serves as a step towards better integration user-centered studies and system-centered studies in IR and also contributes as the theoretical foundation for future researches in this multi-disciplinary topic.

7 REFERENCES

[1] Xu, Y. and Chen, Z. (2006), "Relevance Judgment – What Do Information Users Consider beyond Topicality?", *Journal of the American Society for Information Science and Technology*, 57(7), 961 – 973.

[2] Zhang, Y., Callan, J., and Minka, T. (2002), "Novelty and redundancy detection in adaptive filtering", In *the Proceedings of the 25th Annual International ACM SIGIR Conference on Research and* Development *in Information Retrieval*, Tampere, Finland, August 11-15, 2002. 81-88.

[3] Xu, Y., Yin, H. (2006), "Novelty and Topicality in Information Retrieval", Working paper. National University of Singapore, Singapore; 2006.