The Beauty Contest Revisited: Measuring Consensus Rankings of Relevance using a Game

Christopher G. Harris
Computer Science Department
SUNY Oswego
Oswego, NY 13126
christopher.harris@oswego.edu

ABSTRACT

In this paper, we examine the Keynesian Beauty Contest, a well-known examination of rational agents used to explain the role of consensus predictions in decision making such as price fluctuations in equity markets. Using a game, we study the crowd's ability to judge relevance for both images and textual documents. In addition to asking participants to determine if a document is relevant, we also ask them to rank all choices. One group of participants (N=137) was asked to make judgments based on their own assessment while another group of participants (N = 137) was asked to make judgments based on their estimate of a consensus decision. In addition to measuring recall and precision, our game also uses rank-biased overlap (RBO) to compare each participant's ranked list with the overall consensus decision. Results show the group asked to make ranking decisions based on their estimate of consensus had significantly higher recall for judging relevance in text documents and significantly higher recall and precision when judging relevance for a set of images. We believe this has implications for the determination of consensus across multiple contexts.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Relevance Feedback; Selection Process; H.1.2 [**Models and Principles**]: User / Machine Systems—Human Factors; H.1.2 [**Models and Principles**]: User / Machine Systems—Human information processing; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces

General Terms

Measurement, Economics, Experimentation, Human Factors

Keywords

Gamification, Information Retrieval, Retrieval Effectiveness, Economic Theory, Consensus Prediction, Rank Biased Overlap

1. INTRODUCTION

In 1936, British economist John Maynard Keynes explained the action of rational agents in equity markets using the analogy of a newspaper beauty contest [9]. It was common in this era for British newspapers to ask readers to select the most beautiful woman from a set of six photographs; those choosing the photo with the most votes were entered into a drawing for a prize. The simplest strategy would be for the entrant to select the photo that they personally believed was most beautiful. A more sophisticated strategy would be for the entrant to choose the photo they thought most others would find most beautiful, regardless of the entrant's own personal views. Taking this further, an even more sophisticated strategy might have the entrant modify their choice further based on the viewpoint that other entrants would also alter their selection in anticipation of competitors' strategies.

According to Keynes, this illustrates the rational agent behavior in equity markets, where pricing decisions on a given stock are made not based on what investors' own personal valuations are, but on what they estimate the average of other investors' valuations are likely to be. We examine this same behavior in information retrieval. If, given the results of a query, we ask users to rank the returned documents based on their own perception of relevance, they may order the set one way; ask them to rank these same documents based on a consensus viewpoint, an entirely different ranking may be chosen. In this paper, we examine some of the factors that influence the ability to predict consensus rankings of results. If a ranking of consensus opinions can be predicted with an acceptable level of certainty or to a certain depth, there may be implications in many areas where predicting consensus is important, such as the selection of hot equities, trends in the marketplace, or in the selection of an advertising campaign with the greatest reach.

The contributions of this paper are as follows. First, using a game designed specifically for this purpose, we study how the crowd is able to rank query results against their own perception of what the consensus decision might be. Other research has examined the wisdom of crowds to select a single "best:" choice; however, this is believed to be the first study to examine the relative order of available choices in determining relevance. Second, we examine user decisions on two distinct types of data – one a set of text documents, another a set of images – and evaluate if the ability to determine consensus is substantially different between them.

2. BACKGROUND AND MOTIVATION

Collective decision making has been a focus of the social sciences at least since Plato's *The Republic* (c. 360 BC). In recent years, this discussion has focused on the merits of the wisdom-of-crowds hypothesis, which holds that the independent judgments of a crowd of individuals (as measured by any form of central tendency) will be relatively accurate, even when most of the individuals in the crowd are ignorant and prone to error [15]. The

GamifIR '14, April 13 2014, Amsterdam, Netherlands
Copyright 2014 ACM 978-1-4503-2892-0/14/04 $15.00
http://dx.doi.org/10.1145/2594776.2594780

hypothesis is derived from the following theory: a crowd's judgment comprises signal-plus-noise, averaging judgments will cancel out the noise and extract the signal [8, 12].

According to Surowiecki, four criteria separate wise crowds from irrational ones: diversity of opinion, independence, decentralization, and aggregation. The design of our study meets all of these criteria: each participant has private information (i.e., it is based on their each participants own knowledge or past experiences), maintains independence (i.e., participants are unaware of the decisions made by other participants until after their own decision had been recorded), is decentralized (i.e., participants were from 43 different countries and thus could draw on local knowledge) and last, the decisions of all participants were aggregated based on fair ranking measures of precision, recall, and RBO.

Comparing the quality of the judgments made by the wisdom-of-crowds with those made by experts has become an active area of information retrieval (IR) research. Studies by Alonzo and Mizzaro [1, 2] have examined the quality of crowd workers hired from Amazon Mechanical Turk with those of TREC assessors finding assessment quality was comparable between the two, particularly when a majority decision was taken. Kittur et. al. [10] also found that by redesigning a relevance judgment task to incorporate quality controls the output quality was promising. There have been very few examples using games to make relevance judgments based on consensus. Eickhoff et. al. [4] compared the results generated by a game finding the participants provided high quality outputs. However, participants were not asked to rank documents or images based on their relevance to a given query in any of these tasks.

Learning to Rank (LETOR) is an approach that has gained traction in IR campaigns like TREC, especially with used on datasets with well-known features [11]. However, LETOR requires some advanced knowledge of the document feature set, it prefers a gold standard to train from and it does not work well for predicting subjective decisions, particularly those which contain an unknown prior, making it a poor choice for predicting consensus.

Bayesian Truth Serum (BTS), a concept introduced by Prelec, is a prediction approach for finding truth in biased consensus opinions [13], and can be useful when an established gold standard is not known. BTS improves accuracy by assigning high scores to answers from the crowd that are more common than collectively predicted, with predictions drawn from the same population. It appears to work well with subjective data: Shaw et. al. found that BTS, which involved asking users to estimate the response of their peers, was the best performing of the 14 incentives studied for an online labor task [14]. The BTS approach has a few caveats: the participants need to be Bayesians with a common prior, be sufficiently large enough sample to guarantee truthfulness, and must have the proper incentives to maximizing their score. The game we use in our study follows the latter two but has the benefit in that it does not rely on a common prior.

3. EXPERIMENT METHODOLOGY

3.1 Objective

Our goal is to examine if the crowd makes different decisions about relevance when asked to rank a set of documents based on two mutually-exclusive conditions: (1) based on what they believe most others would decide and (2) based on their own judgment. We look at this with text documents and images. We hypothesize there is a significant difference between the two conditions.

3.1.1 Collections

In our study, we used the following collections.

3.1.1.1 Text Documents

For text we randomly selected 20 query topics from the TREC 8 ad hoc track (topic IDs 401, 403, 405, 406, 407, 411, 413, 417, 418, 419, 420, 421, 422, 424, 425, 427, 429, 436, 437, and 440) [16]. We then randomly selected 5 documents (100 total) from each topic; 50 were judged relevant by the TREC 8 assessors, 50 were documents appearing in at least two-thirds of all 129 submitted TREC 8 ad hoc runs for that topic, but had been pooled and judged as non-relevant by TREC assessors. This provides us with documents that contained keywords that might be construed by many participants as relevant.

3.1.1.2 Image Collections

For images we randomly selected 20 topics from ImageCLEFphoto 2007 track (topic IDs 2, 3, 6, 10, 11, 12, 17, 20, 22, 24, 25, 28, 29, 30, 33, 38, 43, 46, 47, and 52) [5]. We then selected 5 of the corresponding images for each topic (100 total) from a subset of the IAPR TC-12 Benchmark [6]. Fifty of these images were judged relevant by the ImageClef assessors. Although these topics were judged on a ternary classification scheme, we followed the approach in [2] and considered both relevant and partially relevant documents as relevant. Only the image was presented to the participant; none of the provided image annotations were used.

3.2 Participants

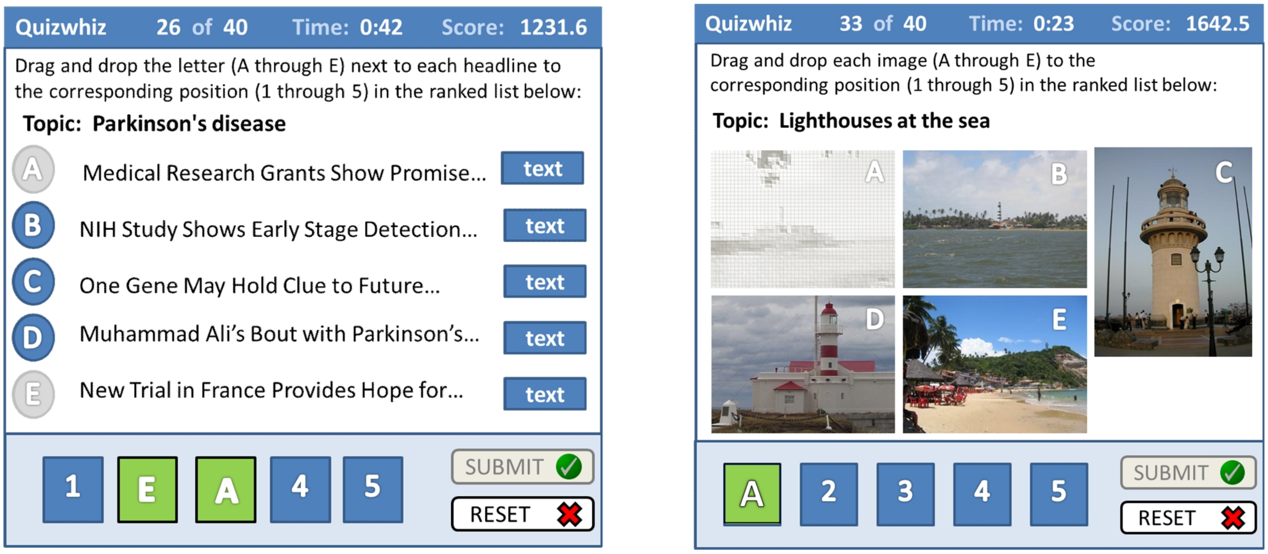
A total of 274 participants from 43 countries were hired from Amazon Mechanical Turk in late December and early January 2014 and given an external URL to participate. Each was paid $0.20 to participate in our game and told that if they obtained one of the 10 highest scores at the end of the 21-day campaign, they would be entered into a drawing for a $5, $10 or $20 prize (which was paid on January 27, 2014). Each participant was told they could only play the game once; we monitored their IP address and their MTurk ID to enforce this restriction. Forty-seven participants (17%) did not complete the entire task or were repeat players; their selections were removed from the consensus decisions and the task relisted for other participants.

3.3 Interface Design and Scoring

All participants were given the same 40 queries and the same 5 images or 5 text documents to rank using a game interface (see Figure 1). To avoid selection bias, the order of questions and answer choices were randomized for each participant. Participants were randomly divided into two groups (Group A and Group B) and given two different sets of instructions. Group A was given the following instructions:

"Rank the following 5 documents (images), from most relevant to least relevant, *based on your own opinion*. Rank these 5 based the information provided in the following query."

Group B is given similar instructions, but told to rank them *based on what they believe most others participating in this task would choose*.



Drag and drop the letter (A through E) next to each headline to the corresponding position (1 through 5) in the ranked list below:

Topic: Parkinson's disease

A Medical Research Grants Show Promise…

B NIH Study Shows Early Stage Detection…

C One Gene May Hold Clue to Future…

D Muhammad Ali's Bout with Parkinson's…

E New Trial in France Provides Hope for…

Drag and drop each image (A through E) to the corresponding position (1 through 5) in the ranked list below:

Topic: Lighthouses at the sea

Figure 1: Screenshot of the game showing text documents (left) and images (right). The letters corresponding to the items already selected and ranked by the participant are shown in the bottom.

The game was designed using PHP and HTML5/CSS3 providing a cohesive experience with a comfortable look and feel similar to [7]. Participants were initially provided with detailed instructions and a complete explanation of the components used in scoring.

3.3.1 Scoring

For game scoring, we used the rank-biased overlap (RBO) method introduced by Webber et. al. [17]. RBO, being a set-based measure has an advantage over rank-based measures, such as Kendalls' Tau since RBO is top-weighted (i.e., documents at the top of the ranked list count more towards the game score) and is easy to calculate. For example, given the following rank for user X = {a, b, c, d, e} with the consensus rank Y = {b, a, c, d, e}, Table 1 provides an overview of RBO scoring for X and Y:

Table 1. An example showing a mismatch at the top of the ranked list and the scoring obtained through RBO.

Depth (k)	Items in List X @ k	Items in List Y @ k	Set intersection	Fraction
1	a	b	{}	0/2 = 0
2	a,b	b,a	{a,b}	2/2 = 1
3	a,b,c	b,a,c	{a,b,c}	3/3 = 1
4	a,b,c,d	b,a,c,d	{a,b,c,d}	4/4 = 1
5	a,b,c,d,e	b,a,c,d,e	{a,b,c,d,e}	5/5 = 1

Table 2. An example showing a mismatch near the bottom of the ranked list and the scoring obtained through RBO.

Depth (k)	Items in List X @ k	Items in List Z @ k	Set intersection	Fraction
1	a	a	{a}	1/1 = 1
2	a,b	a,b	{a,b}	2/2 = 1
3	a,b,c	a,b,c	{a,b,c}	3/3 = 1
4	a,b,c,d	b,a,c,e	{a,b,c}	3/5 = 0.6
5	a,b,c,d,e	b,a,c,e,d	{a,b,c,d,e}	5/5 = 1

Using RBO score, the results are (0+1+1+1+1)/(1+1+1+1+1)= 4/5 = 0.8 (see Table 1), the same as a rank-based measure. If a second participant, Z, ranked these items as {a, b, c, e, d}, a rank-based measure would give participant Z a score of 0.8 as well; however RBO scores this as (1+1+1+1+0.6+1)/(1+1+1+1+1) = 4.6/5 = 0.92 (see Table 2), reflecting greater uniformity at the top [3]. We multiplied this RBO score by 100 to give a base score in the range (0, 100) for each question.

3.3.2 Bonus

We also rewarded participants with a bonus based on the time required to complete each question. Participants taking longer than the median time for a question are not given a bonus (we use median instead of mean as the median is less sensitive to large outliers). The calculation was a fraction of their base score for that question multiplied by 0.2 x the standard deviation from the median time taken to complete that same question; e.g., if a question took a participant 70 seconds to complete and the median time taken by all others was 100 seconds with a standard deviation of 15 seconds, they would obtain a bonus of 0.2 x (100-70)/15 = 0.4 times their base score for that question. All participants, including those in Group A (the group asked to rank based on their own opinion) were also told their final game score could change before the end of the campaign.

3.3.3 Precision and Recall

In addition, participants were also asked to mark which of the choices for each topic they believed were relevant (for Group A) or that they believed consensus would determine to be relevant (for Group B); therefore, we also evaluated precision and recall for each participant against the gold standard. However, recall and precision were not components of the game score.

3.3.4 Gamification Design Issues

Using a game format to determine consensus, we need to be concerned about several issues, including bias, cheating and the cold-start problem. Selection bias is addressed by randomizing questions and the items. Moreover, bias in completion times for game scoring is addressed by using the median, instead of the mean, time. We addressed potential cheating issues by monitoring IP addresses and the MTurk IDs of participants.

A cold start problem occurs in systems that rely on inferences (as we do for a consensus decision) but have not yet gathered sufficient information to obtain a consensus. For the first 10 participants, we score their rankings randomly; for subsequent participants, we use information from all previous decisions.

Two separate leaderboards (one for each group) was provided to participants at the beginning of the game and at the end of the game. Participants were provided with a URL and could view the leaderboard until the campaign completed on January 26, 2014. Bonuses were paid to participants on the leaderboard at the end of the campaign.

4. RESULTS

In order to examine the effectiveness of individual vs. consensus decisions, we examined the difference between Group A (instructed to provide their own opinion) and Group B (instructed to provide a consensus opinion) for text, for images, and overall across each of four performance measures: precision, recall, F-score, and RBO. These metrics are provided in Table 3. We discuss each performance measure and provide additional details separately in sections 4.1 through 4.3.

Table 3. Performance metrics for both groups, broken out by text and images. (*= statistically significant improvement)

Type of Query	Precision	Recall	RBO
Group A (individual opinion)			
Text	0.5179	0.4922	0.5504
Images	0.5874	0.6036	0.6321
Overall	0.5527	0.5479	0.5913
Group B (consensus opinion)			
Text	0.5348	0.5108*	0.5988*
Images	0.6745*	0.7294*	0.7024*
Overall	0.6047*	0.6208*	0.6506*

4.1 Precision

Using a two-tailed, independent sample t-test, we examined the difference in precision between the two groups at $p < 0.05$. For queries on text documents, there was not a significant difference between the two groups in precision, σ_A=0.095, σ_B=0.078, t(272) = 1.61, p = 0.1084. For queries on images, Group B significantly outperformed Group A on precision, σ_A=0.085, σ_B=0.091, t(272) = 8.19, s= $p < 0.001$. Taking both types of queries together, Group B significantly outperformed Group A on precision σ_A=0.090, σ_B=0.084, t(272) = 4.93, $p < 0.001$.

4.2 Recall

Using a two-tailed, independent sample t-test, we examined the difference in recall between the two groups at $p < 0.05$. For queries on text documents, Group B significantly outperformed Group A in recall, σ_A=0.076, σ_B=0.081, t(272) = 2.10, p = 0.037. Likewise, for queries on images, Group B significantly outperformed Group A in recall, σ_A=0.082, σ_B=0.085, t(272) = 12.52, $p < 0.001$. Taking both types of queries together, Group B significantly outperformed Group A on recall, σ_A=0.080, σ_B=0.083, t(272) = 7.46, $p < 0.001$.

4.3 Rank-Biased Overlap (RBO)

Using a two-tailed, independent sample t-test, we examined the difference in RBO between the two groups at $p < 0.05$. For queries on text documents, Group B significantly outperformed Group A in RBO, σ_A=0.065, σ_B=0.070, t(272) = 2.10, p = 0.037. Likewise, for queries on images, Group B significantly outperformed Group A in RBO, σ_A=0.069, σ_B=0.070, t(272) = 12.52, $p < 0.001$. Taking both types of queries together, Group B significantly outperformed Group A on RBO, σ_A=0.067, σ_B=0.070, t(272) = 7.46, $p < 0.001$.

5. ANALYSIS

From these observations, we determine that the group making decisions based on consensus (Group B) significantly outperformed the group making decisions based on their own opinions (Group A) on all three performance measures. For text, Group B outperformed Group A in recall and RBO, and for images, Group B outperformed Group A in all performance measures. Thus, we find that when game participants are instructed to rank documents based on consensus opinion, it enhances their ability in precision, recall, and RBO for images and for recall and RBO for text documents.

Overall, the game participants in Group A had a much larger standard deviation for each of our performance measures as compared with those in Group B. This may indicate that when participants are asked to use consensus opinion as a guide, they are more conservative in their determination of relevance. This may explain some of the dynamics of group-based decision making in other contexts: people may be more reluctant to provide diverse opinions or embrace other viewpoints of relevance that may be "out of the box".

We also examined each of the 40 topics in more detail. We observed several topics in which the difference in the determination of relevance between the two groups was very large (e.g., text topic IDs 421 and 427 as well as image topic IDs 11, 20, and 43). Participants asked to make decisions based on consensus were more cautious in marking documents as relevant, which negatively affected recall on these topics. Surprisingly, the rank order (and RBO score) for these same topics were not significantly affected. This may indicate the most appropriate ranking of documents or images could be ascertained even when the appropriate threshold of relevance could not. Additionally, using a ranking scheme instead of a binary relevance decision is likely more effective as it provides a more holistic picture of how people actually perceive relevance. This may explain how decisions that do not meet specific conditions are made across different contexts. We plan to explore how people perceive relevance in future research.

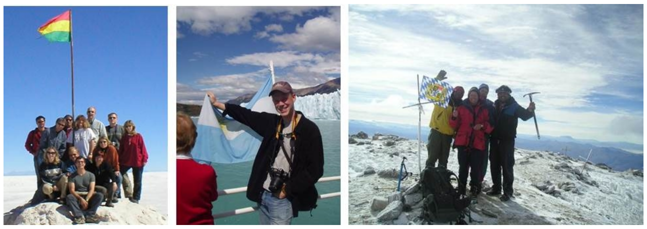


Figure 2: Three of the images presented for Topic ID 25 "people with a flag". Since all three are relevant, one challenge to participants is how to properly rank them.

When multiple images or text documents meet the criteria stated in the information need, it can be challenging to rank them, and far more challenging to rank them in a consistent manner. Consider the images in Figure 2 – all three meet the criteria of Topic ID 25 "People with a flag", but we found the participants asked to rank based on consensus ranked them based on the prominence of the objects stated in the information need whereas those asked to rank them based on their own opinions were far less consistent with a ranking order, and less consistent with the consensus rank as well. We noticed this pattern across topics for both text and images. This implies a different level of thinking for the consensus participants, which is consistent with the Keynesian Beauty Contest.

6. CONCLUSION

In this paper, we have examined the principles of the Keynesian Beauty Contest, an analogy once used to explain how people make rational decisions based on their expectation of how most others will behave in a given situation. Our study focused on the instructions provided to participants in a game – one group was told to rank a set of text documents or images based their own opinion whereas another group was told to rank based on their expectation of the overall consensus opinion. Our game provided incentives to have participants quickly rank five items based on a provided information need.

We found that participants ranking based on their expectation of consensus outperformed those who ranked based on their own opinions on all three performance measures. The increase was more significant for images than text. Those provided a consensus opinion were able to consistently rank items, even when they were all considered relevant to the information need. In images, the primary consideration was the main action within the photo, and with text documents it was the percent of the article that discussed the information need. We uncovered caveats to using consensus prediction, however, such as less divergent thinking, which we plan to examine in future work.

7. REFERENCES

[1] Alonso, O., & Mizzaro, S. (2009, July). Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation (pp. 15-16).

[2] Alonso, O., & Mizzaro, S. (2012). Using crowdsourcing for TREC relevance assessment. Information Processing & Management, 48(6), 1053-1066.

[3] Argawal, R. Comparing Ranked List. 2013. Retrieved January 19, 2014 from http://ragrawal.wordpress.com/2013/01/18/comparing-ranked-list/

[4] Eickhoff, C., Harris, C. G., de Vries, A. P., & Srinivasan, P. (2012, August). Quality through flow and immersion: gamifying crowdsourced relevance assessments. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (pp. 871-880). ACM.

[5] Grubinger, M., Clough, P., Hanbury, A., & Müller, H. (2008). Overview of the ImageCLEFphoto 2007 photographic retrieval task. In Advances in Multilingual and Multimodal Information Retrieval (pp. 433-444). Springer Berlin Heidelberg.

[6] Grubinger, M., Clough, P., Müller, H., & Deselaers, T. (2006). The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In International Workshop OntoImage (pp. 13-23).

[7] Harris, C. G., & Srinivasan, P. (2013). Comparing crowd-based, game-based, and machine-based approaches in initial query and query refinement tasks. In Advances in Information Retrieval (pp. 495-506). Springer Berlin Heidelberg.

[8] Hogarth, R. M. (1978). A note on aggregating opinions. Organizational Behavior and Human Performance, 21(1), 40-46.

[9] Keynes, J. M. (1936). General theory of employment, interest and money. Atlantic Books.

[10] Kittur, A., Chi, E. H., & Suh, B. (2008, April). Crowdsourcing user studies with Mechanical Turk. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 453-456). ACM.

[11] Liu, T. Y. (2009). Learning to rank for information retrieval. Foundations and Trends in Information Retrieval, 3(3), 225-331.

[12] Makridakis, S., & Winkler, R. L. (1983). Averages of forecasts: Some empirical results. Management Science, 29(9), 987-996.

[13] Prelec, D. (2004). A Bayesian truth serum for subjective data. Science, 306(5695), 462-466.

[14] Shaw, A. D., Horton, J. J., & Chen, D. L. (2011, March). Designing incentives for inexpert human raters. In Proceedings of the ACM 2011 conference on Computer supported cooperative work (pp. 275-284). ACM.

[15] Surowiecki, J. (2005). The wisdom of crowds. Random House Digital, Inc..

[16] Vorhees, V., & Harman, D. (1999). Overview of the eighth Text REtrieval Conference (TREC-8). In Proc. TREC.

[17] Webber, W., Moffat, A., & Zobel, J. (2010). A similarity measure for indefinite rankings. ACM Transactions on Information Systems (TOIS), 28(4), 20.