Stochastic Relevance for Crowdsourcing

Marco Ferrante[1], Nicola Ferro[2], and Eleonora Losiouk[1]

[1] Department of Mathematics "Tullio Levi-Civita", University of Padua, Italy
{ferrante, elosiouk}@math.unipd.it
[2] Department of Information Engineering, University of Padua, Italy
ferro@dei.unipd.it

Abstract. It has been recently proposed to consider relevance assessment as a stochastic process where relevance judgements are modeled as binomial random variables and, consequently, evaluation measures become random evaluation measures, removing the distinction between binary and multi-graded evaluation measures.

In this paper, we adopt this stochastic view of relevance judgments and we investigate how this can be applied in the crowd-sourcing context. In particular, we show that injecting some randomness in the judgments by crowd assessors improves their correlation with the gold standard and we introduce a new merging approach, based on binomial random variables, which is competitive with respect to state-of-the-art at low numbers of merged assessors.

1 Introduction

It has been recently proposed to model relevance assessment as a stochastic process where each relevance judgement is a binomial random variable whose expectation p indicates the quantity of relevance assigned to a document [5]. This choice allowed for seamlessly modeling both binary and graded relevance judgements into a single framework and for introducing the notion of *random evaluation measures*, which are just a transformation of such binomial variables, eliminating the distinction between binary and graded evaluation measures.

In this paper, we investigate to what extent this new way of modelling relevance judgements can be applied in the context of crowdsourcing [1]. In particular, we study the following research questions:

RQ1 how the random evaluation measures can improve the robustness to variations in the assessments;

RQ2 how the proposed binomial framework can be extended to allow for merging multiple crowd-assessors.

We conduct a systematic experimentation using the TREC 2012 Crowdsourcing track [13] in order to answer the two research questions above.

The paper is organized as follows: Section 2 discusses some related works; Section 3 introduces our stochastic framework for merging crowd assessors; Section 4 reports the evaluation results; and, Section 5 draws some conclusions and outlooks possible future works.

2 Related Work

2.1 Crowdsourcing

Crowdsourcing [1, 9–11] has emerged as a viable option for ground-truth creation since it allows to cheaply collect multiple assessments for each document. However, it raises many questions regarding the quality of the collected assessments. Therefore, in order to obtain a ground-truth good enough to be used for evaluation purposes, the possibility of discarding the low quality assessors and/or combining them with more or less sophisticated algorithms has been considered.

State of the art crowdsourcing algorithms are *Majority Vote*, where the label with the highest number of votes, i.e. assessors, is selected, and *Expectation-Maximization* [2, 6], where the Expectation-Maximization algorithm is used to iteratively select the most probable labels. More recently, AWARE [4] has been proposed as a way to compute a weighted mean of evaluation measures computed for each crowd assessor.

2.2 Binomial Relevance Framework

Ferrante et al. [5] described the relevance of a document via a **binomial random variable** $Bi(1, p)$ with parameters 1 and p, where p roughly defines the *quantity of relevance* of that document. For each topic, document pair $(t, d_i) \in T \times D$, they defined the **random ground-truth** RGT, also called random relevance, as a binomial random variable of parameters $(1, p_{t,d_i})$, where p_{t,d_i} is the quantity of relevance associated to the document d_i with respect to a topic t. In this framework, $p_{t,d_i} = 0$ corresponds to a document completely not relevant and $p_{t,d_i} = 1$ to a fully relevant document.

Thanks to the random ground-truth, they turned every evaluation measure into a random evaluation measure, by simply composing the original expression of each measure with the random relevances. To compare different systems, they needed to define an ordering among runs and, to this end, they used the expected values of the random measures defined above.

Therefore, **expected Random Rank Biased Precision (eRRBP)** is

$$\mathbb{E}\big[RBP[\hat{r}_t(\omega)]\big] = (1-\tau)\sum_{n=1}^{N}\tau^{n-1}p_{t,d_n}$$

where τ represents the persistence.

Then, **expected Random Discounted Cumulative Gain (eRDCG)** is

$$\mathbb{E}\big[DCG[\hat{r}_t(\omega)]\big] = \sum_{n=1}^{N}\frac{p_{t,d_n}}{\max\{1, \log_{10}(n)\}}$$

Finally, **expected Random Average Precision (eRAP)** is

$$\mathbb{E}\big[AP[\hat{r}_t(\omega)]\big] = \frac{1}{\widehat{RB}_t}\sum_{n=1}^{N}\frac{1}{n}\left(1+\sum_{s=1}^{n-1}p_{t,d_s}\right)p_{t,d_n}$$

where $\widehat{RB}_t = \sum_{d \in D}\mathbb{E}\big[RGT(t,d)\big]$ is the expected recall base.

3 Random Relevance for Merging Crowd-Assessors

Let us assume that M assessors evaluate a pool of documents $\{d_1, \dots, d_N\}$ with respect to a topic $t \in T$. According to [5], the judgment of the j-th assessor for the pair (t, d_i) is a Binomial random variable $AS_j(t, d_i)$ which models the amount of relevance of the document according to that assessor.

We assume that for any pair (t, d_i), $AS_1(t, d_i), \dots, AS_M(t, d_i)$ are independent binomial random variables of parameters $(1, p_{t,d_i})$. Note that this i.i.d. assumption is implicitly done in all the previous works about merging crowd-assessors.

We leverage the assumption above to define the **Binomial Majority Vote (BINMV)** merging strategy, where the unknown parameter p_{t,d_i}, i.e. the merged amount of relevance of each topic/document pair, is estimated from the observed values of the random variables $AS_1(t, d_i), \dots, AS_M(t, d_i)$ as

$$\widetilde{p}_{t,d_i} = \frac{1}{M} \sum_{j=1}^{M} AS_j(t, d_i)$$

and we define the random ground-truth $RGT(t, d_i)$ as a Binomial random variable of parameters $(1, \widetilde{p}_{t,d_i})$. As the name suggests, this strategy adopts the same logic as the Majority Vote approach but applied in the case of the random relevance.

We also define the **Quantized Binomial Majority Vote (QBINMV)** strategy which applies a sigmoid function $\frac{1}{1+\exp^{-k*(x-0.5)}}$ to the estimated parameter $\widetilde{p}_{t,d_i}$ in order to reduce the number of relevance degrees produced by the BINMV strategy and make sharper decisions towards being relevant or not relevant; in particular, we use $k = 15$.

4 Experiments

4.1 Experimental Setup

We use the TREC 21, 2012, Crowdsourcing (`T21`) [13] data set developed in the *Text Relevance Assessing Task (TRAT)*. The TRAT required participating groups to simulate the relevance assessing role of the NIST for 10 of the TREC 08, 1999, Ad-hoc topics [16], using binary relevance. In total 33 pools were submitted to TRAT; we excluded two of them (`INFLB2012` and `Orc2Stage`) because, for some topics, they did not assess any document as relevant.

Two TREC Adhoc tracks used these 10 topics over the years: the TREC 08, 1999, Ad-hoc track [16] (labeled `T08`), which contains 129 runs; and, the TREC 13, 2004, Robust track [15] (labeled `T13`), which contains 110 runs.

As in the TREC crowdsourcing track, we use correlation analysis – both Kendall's τ correlation [8] and AP correlation τ_{AP} [18] – to compare crowd assessors with respect to the gold standard pool.

We consider the following evaluation measures, to be compared against their random version: AP [3], DCG [7], and RBP [12]. We use log base 10 for DCG

Table 1. τ and τ_{AP} averaged over the T21 crowd-assessors. Gold standard is labelled as: DGS (Deterministic Gold Standard); RGS (Randomized Gold Standard).

		T08 Systems		T13 Systems	
		Mean τ	**Mean τ_{AP}**	**Mean τ**	**Mean τ_{AP}**
DGS	**AP**	0.7023 ± 0.0522	0.5802 ± 0.0655	0.7044 ± 0.0532	0.5655 ± 0.0679
DGS	**eRAP**	0.6704 ± 0.0436	0.5437 ± 0.0485	0.7033 ± 0.0355	0.5551 ± 0.0444
RGS	**eRAP**	**0.7077 ± 0.0537**	**0.5900 ± 0.0608**	**0.7471 ± 0.0469**	**0.6056 ± 0.0642**
DGS	**DCG**	0.7222 ± 0.0454	0.5998 ± 0.0482	0.7621 ± 0.0466	0.6161 ± 0.0601
DGS	**eRDCG**	0.6896 ± 0.0373	0.5737 ± 0.0422	0.7391 ± 0.0433	0.5833 ± 0.0585
RGS	**eRDCG**	**0.7858 ± 0.0356**	**0.6776 ± 0.0436**	**0.7766 ± 0.0396**	**0.6240 ± 0.0522**
DGS	**RBP**	0.6732 ± 0.0547	0.5341 ± 0.0684	0.5879 ± 0.0657	0.4534 ± 0.0706
DGS	**eRRBP**	0.6739 ± 0.0546	0.5352 ± 0.0684	0.5904 ± 0.0647	**0.4560 ± 0.0699**
RGS	**eRRBP**	**0.6749 ± 0.0545**	**0.5359 ± 0.0683**	**0.5922 ± 0.0645**	0.4555 ± 0.0698

and gains 0 and 1 for not relevant and relevant documents, respectively; we use persistence $p = 0.8$ for RBP.

To ease the reproducibility of the experiments, the source code is available at: `https://bitbucket.org/frrncl/ecir2019-ffl/`.

4.2 RQ1: Robustness to Variations in the Assessments

For each crowd-assessor submitted to the T21 track, we computed the τ and τ_{AP} correlations with respect to the gold standard pool and then we averaged these scores over all the crowd-assessors. Table 1 reports the summary averages together with their confidence intervals. For each measure, we report: (i) the state-of-the-art deterministic version compared against a Deterministic Gold Standard (DGS); the random version using $p_{notrel} = 0.05$ and $p_{rel} = 0.95$ for the crowd-assessors, i.e. we allow for just a small 5% confidence on their judgements, compared against the DGS, i.e. the same used for the deterministic measures; the random version as before but compared against a Randomized Gold Standard (RGS), which is the gold standard pool but using $p_{notrel} = 0.05$ and $p_{rel} = 0.95$, i.e. we assume just a small randomness also in it.

Comparing the deterministic measures against DGS to the random ones against RGS, we can observe how the random evaluation measures substantially improve the average agreement among the gold standard and the crowd-assessors, consistently for both τ and τ_{AP} and across both tracks, T08 and T13.

Comparing the deterministic measures against DGS to the random ones against DGS, we can observe that deterministic measures tend to perform better, with the exception of RBP and eRBP whose performance are almost the same. However, it should be noted that this is by far the most unfavourable comparison for the random evaluation measures, since the DGS pool does not account for any kind of randomness and awards only the deterministic evaluation measures.

Overall, we can conclude that injecting some randomness into the evaluation measures is beneficial for compensating variations in relevance judgements in a crowdsourcing context.

Table 1 also opens an important question about what we should consider as gold standard: a deterministic or a random pool? If, for example, we consider the inter-assessor agreement issue [14, 17], we should conclude that the gold standard we daily use in evaluation campaigns is far from being deterministic and, perhaps, we should move to a stochastic vision of it.

Table 2. τ_{AP} for different merging strategies and different numbers M of merged assessors using `T08` and `T13` systems. Gold standard is labelled as: DGS (Deterministic Gold Standard); RGS (Randomized Gold Standard).

	T08 **Systems**							
		$M = 2$	$M = 3$	$M = 4$	$M = 5$	$M = 10$	$M = 20$	$M = 30$
DGS	**MV AP**	0.5757	0.6425	0.7135	0.6920	**0.7605**	**0.7979**	**0.8103**
	EM AP	0.5722	0.6161	0.7147	0.6749	0.7445	0.7522	0.7443
	AWARE AP	**0.6797**	0.6525	0.7124	**0.6928**	0.7138	0.7034	0.7089
DGS	**BINMV eRAP**	0.5924	0.5807	0.6050	0.5978	0.5794	0.5617	0.5657
	QBINMV eRAP	0.5921	0.6299	0.7028	0.6696	0.6913	0.6848	0.6615
RGS	**BINMV eRAP**	0.6037	0.6173	0.6432	0.6259	0.6329	0.6211	0.6225
	QBINMV eRAP	0.6043	**0.6561**	**0.7226**	0.6663	0.7441	0.7445	0.7314
DGS	**MV DCG**	0.6123	0.6901	0.7116	0.6733	0.7432	**0.7868**	**0.7895**
	EM DCG	0.5441	0.6741	0.7014	0.6642	0.6770	0.6756	0.6598
	AWARE DCG	0.6190	0.6397	0.6540	0.6392	0.6461	0.6499	0.6508
DGS	**BINMV eRDCG**	0.6190	0.6397	0.6540	0.6392	0.6461	0.6499	0.6508
	QBINMV eRDCG	0.6187	0.6830	0.7096	0.6735	0.7286	0.7453	0.7544
RGS	**BINMV eRDCG**	0.6878	0.7181	0.7371	**0.7265**	0.7507	0.7502	0.7559
	QBINMV eRDCG	**0.6884**	**0.7410**	**0.7517**	0.7185	**0.7660**	0.7794	0.7892
DGS	**MV RBP**	0.6211	0.6138	0.7065	0.6661	0.7243	0.7444	0.7365
	EM RBP	0.5194	0.6087	0.6937	0.6214	0.6790	0.7053	0.6780
	AWARE RBP	0.6421	0.6374	0.7109	0.6811	0.7139	0.7152	0.7205
DGS	**BINMV eRRBP**	0.6422	0.6374	0.7109	0.6811	0.7139	0.7152	0.7205
	QBINMV eRRBP	0.6422	0.6277	**0.7241**	0.6835	**0.7453**	0.7602	**0.7716**
RGS	**BINMV eRRBP**	0.6428	**0.6381**	0.7114	0.6817	0.7147	0.7160	0.7213
	QBINMV eRRBP	**0.6429**	0.6279	0.7240	**0.6836**	0.7451	**0.7603**	**0.7716**

	T13 **Systems**							
		$M = 2$	$M = 3$	$M = 4$	$M = 5$	$M = 10$	$M = 20$	$M = 30$
DGS	**MV AP**	0.6158	0.6253	0.7167	0.7138	0.7674	**0.7995**	**0.8226**
	EM AP	0.5486	0.5575	0.7218	0.6877	0.7441	0.7833	0.7704
	AWARE AP	**0.6965**	**0.6717**	**0.7477**	**0.7221**	0.7693	0.7591	0.7600
DGS	**BINMV eRAP**	0.6315	0.6135	0.6648	0.6553	0.6485	0.6538	0.6513
	QBINMV eRAP	0.6306	0.6443	0.7106	0.7033	0.7248	0.7308	0.7248
RGS	**BINMV eRAP**	0.6381	0.6451	0.7130	0.6798	0.7271	0.7116	0.7073
	QBINMV eRAP	0.6411	0.6696	0.7320	0.6898	**0.7731**	0.7828	0.7849
DGS	**MV DCG**	0.6233	0.7039	**0.7516**	**0.7122**	**0.7923**	**0.8129**	**0.8226**
	EM DCG	0.5719	**0.7044**	0.7496	0.6972	0.7254	0.7211	0.6845
	AWARE DCG	0.6349	0.6496	0.6758	0.6542	0.6616	0.6565	0.6517
DGS	**BINMV eRDCG**	0.6349	0.6496	0.6758	0.6542	0.6616	0.6565	0.6517
	QBINMV eRDCG	0.6347	0.6919	0.7436	0.7062	0.7795	0.7815	0.7774
RGS	**BINMV eRDCG**	0.6466	0.6590	0.6990	0.6749	0.7051	0.6987	0.6975
	QBINMV eRDCG	**0.6470**	0.6938	0.7214	0.6975	0.7455	0.7541	0.7570
DGS	**MV RBP**	0.4893	0.5036	0.5880	0.5385	0.6114	0.6062	0.6187
	EM RBP	0.4106	0.4974	0.5989	0.5180	0.6270	0.6094	0.5943
	AWARE RBP	0.5614	0.5586	0.6125	**0.5717**	0.6342	**0.6307**	**0.6248**
DGS	**BINMV eRRBP**	0.5614	0.5586	0.6125	**0.5717**	0.6342	**0.6307**	**0.6248**
	QBINMV eRRBP	0.5614	0.5233	0.6038	0.5594	0.6292	0.6263	0.5988
RGS	**BINMV eRRBP**	**0.5619**	**0.5589**	**0.6137**	**0.5717**	**0.6348**	0.6304	0.6245
	QBINMV eRRBP	**0.5619**	0.5232	0.6042	0.5592	0.6297	0.6263	0.5990

4.3 RQ2: Random Relevance for Merging Crowd-Assessors

Let $L = 31$ be the total number of available crowd-assessors and $M < L$ the number of assessors we are merging. For each of the above evaluation measures, we experimented all the $M = 2, 3, \ldots, 30$. For each value of M, there are $\binom{L}{M} = \binom{31}{M} = \frac{31!}{M!(31-M)!}$ possible ways of choosing the M assessors to be merged; we randomly sampled 10 M-tuples out of the $\binom{31}{M}$ possible ones.

Table 2 reports the average of τ_{AP} correlation over these 10 samples for both `T08` and `T13` systems; the results using Kendall's τ correlation are similar but not reported here for space reasons. As in the case of Table 1, for each measure, we report: (i) the state-of-the-art deterministic merging strategy compared against DGS; the random merging strategy compared against the DGS; the random merging strategy compared against RGS. As state-of-the-art deterministic merging strategy we considered Majority Vote (MV), Expectation-Maximization (EM), and AWARE with uniform weights.

If we compare the results of Table 2 with those of Table 1 we can note how all the merging strategies improve with respect to the performance of single crowd assessors.

When it comes to merging in the case of the deterministic state-of-the-art merging strategies, we can observe that MV is always the most effective approach for high numbers of merged assessors while AWARE is competitive for lower numbers, a more interesting case due to the less resources required. EM tends to have lower performance when using fewer assessors and they increase for more assessors but almost never reaching MV.

BINMV is especially effective with eRRBP, which always improves for low numbers of assessors with respect to MV and AWARE. However, in the case of eRAP and eRDCG deterministic state-of-the-art merging strategies tend to perform better. However, as discussed in the case of RQ1, the RGS is a more fair comparison for the random evaluation measures and, in this case, we can observe more substantial improvements for the BINMV strategy which often outperforms state-of-the-art ones.

Finally, QBINMV is typically more effective than BINMV and it often performs better than deterministic state-of-the-art merging strategies. This is probably due to the fact that it reduces the number of relevance degrees, which is almost "continuous" in the case of the BINMV strategy, and pushes towards choosing between either relevant or not relevant. This makes QBINMV closer to the deterministic evaluation measures, which use just binary relevance, and so they compete on a closer basis.

5 Conclusions and Future Work

In this paper, we investigated how a stochastic approach for modelling relevance as a random binomial variable behaves in the context of crowdsourcing. We have shown how injecting some randomness in the relevance judgments of crowd-assessors improves their correlation with the gold standard (RQ1). We have also shown how the binomial relevance framework can be used to develop new merging strategies which are competitive with respect to state-of-the-art when using fewer crowd-assessors, which means reducing the required resources (RQ2). In both cases, the conducted investigation raised the issue of whether it is more appropriate to use a deterministic or a randomized gold standard.

Overall, we can appreciate the benefits of moving to a random relevance framework which is capable to unify into a single coherent vision binary to multi-graded relevance, management of incomplete information and variations in relevance judgments, and merging of crowd-assessors.

As future work, we will investigate how using a stochastic gold standard instead of a deterministic one impacts on IR evaluation. Moreover, we plan to leverage the binomial relevance framework to develop more advanced merging strategies able to also account for the quality of the assessors, instead of simply merging them in a uniform way.

References

1. Alonso, O., Mizzaro, S.: Using Crowdsourcing for TREC Relevance Assessment. Information Processing & Management 48(6), 1053–1066 (November 2012)
2. Bashir, M., Anderton, J., Wu, J., Ekstrand-Abueg, M., Golbus, P.B., Pavlu, V., Aslam, J.A.: Northeastern University Runs at the TREC12 Crowdsourcing Track. In: Voorhees, E.M., Buckland, L.P. (eds.) The Twenty-First Text REtrieval Conference Proceedings (TREC 2012). National Institute of Standards and Technology (NIST), Special Publication 500-298, Washington, USA (2013)
3. Buckley, C., Voorhees, E.M.: Retrieval System Evaluation. In: Harman, D.K., Voorhees, E.M. (eds.) TREC. Experiment and Evaluation in Information Retrieval. pp. 53–78. MIT Press, Cambridge (MA), USA (2005)
4. Ferrante, M., Ferro, N., Maistro, M.: AWARE: Exploiting Evaluation Measures to Combine Multiple Assessors. ACM Transactions on Information Systems (TOIS) 36(2), 20:1–20:38 (September 2017)
5. Ferrante, M., Ferro, N., Pontarollo, S.: Modelling Randomness in Relevance Judgments and Evaluation Measures. In: Pasi, G., Piwowarski, B., Azzopardi, L., Hanbury, A. (eds.) Advances in Information Retrieval. Proc. 40th European Conference on IR Research (ECIR 2018). pp. 197–209. Lecture Notes in Computer Science (LNCS) 10772, Springer, Heidelberg, Germany (2018)
6. Hosseini, M., Cox, I.J., Milić-Frayling, N., Kazai, G., Vinay, V.: On Aggregating Labels from Multiple Crowd Workers to Infer Relevance of Documents. In: Baeza-Yaetes, R., de Vries, A.P., Zaragoza, H., Cambazoglu, B.B., Murdock, V., Lempel, R., Silvestri, F. (eds.) Advances in Information Retrieval. Proc. 32nd European Conference on IR Research (ECIR 2012). pp. 182–194. Lecture Notes in Computer Science (LNCS) 7224, Springer, Heidelberg, Germany (2012)
7. Järvelin, K., Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques. ACM Transactions on Information Systems (TOIS) 20(4), 422–446 (October 2002)
8. Kendall, M.G.: Rank correlation methods. Griffin, Oxford, England (1948)
9. King, I., Chen, K.T., Alonso, O., Larson, M.: Special Issue: Crowd in Intelligent Systems. ACM Transactions on Intelligent Systems and Technology (TIST) 7(4) (May 2016)
10. Lease, M., Yilmaz, E.: Crowdsourcing for Information Retrieval: Introduction to the Special Issue. Information Retrieval 16(2), 91–100 (April 2013)
11. Marcus, A., Parameswaran, A.: Crowdsourced Data Management: Industry and Academic Perspectives. Foundations and Trends in Databases (FnTDB) 6(1–2), 1–161 (December 2015)
12. Moffat, A., Zobel, J.: Rank-biased Precision for Measurement of Retrieval Effectiveness. ACM Transactions on Information Systems (TOIS) 27(1), 2:1–2:27 (December 2008)
13. Smucker, M.D., Kazai, G., Lease, M.: Overview of the TREC 2012 Crowdsourcing Track. In: Voorhees, E.M., Buckland, L.P. (eds.) The Twenty-First Text REtrieval Conference Proceedings (TREC 2012). National Institute of Standards and Technology (NIST), Special Publication 500-298, Washington, USA (2013)
14. Voorhees, E.M.: Variations in relevance judgments and the measurement of retrieval effectiveness. In: Croft, W.B., Moffat, A., van Rijsbergen, C.J., Wilkinson, R., Zobel, J. (eds.) Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998). pp. 315–323. ACM Press, New York, USA (1998)

15. Voorhees, E.M.: Overview of the TREC 2004 Robust Track. In: Voorhees, E.M., Buckland, L.P. (eds.) The Thirteenth Text REtrieval Conference Proceedings (TREC 2004). National Institute of Standards and Technology (NIST), Special Publication 500-261, Washington, USA (2004)
16. Voorhees, E.M., Harman, D.K.: Overview of the Eigth Text REtrieval Conference (TREC-8). In: Voorhees, E.M., Harman, D.K. (eds.) The Eighth Text REtrieval Conference (TREC-8). pp. 1–24. National Institute of Standards and Technology (NIST), Special Publication 500-246, Washington, USA (1999)
17. Webber, W., Chandar, P., Carterette, B.A.: Alternative Assessor Disagreement and Retrieval Depth. In: Chen, X., Lebanon, G., Wang, H., Zaki, M.J. (eds.) Proc. 21st International Conference on Information and Knowledge Management (CIKM 2012). pp. 125–134. ACM Press, New York, USA (2012)
18. Yilmaz, E., Aslam, J.A., Robertson, S.E.: A New Rank Correlation Coefficient for Information Retrieval. In: Chua, T.S., Leong, M.K., Oard, D.W., Sebastiani, F. (eds.) Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008). pp. 587–594. ACM Press, New York, USA (2008)