The Impact of Relevance Judgments and Data Fusion on Results of Image Retrieval Test Collections

William Hersh, Eugene Kim

Department of Medical Informatics & Clinical Epidemiology

School of Medicine

Oregon Health & Science University

3181 SW Sam Jackson Park Rd., BICC

Portland, OR, USA 97239

e-mail: hersh@ohsu.edu

Abstract

The goal of this study was to determine how varying relevance judgments impact the absolute and relative performance of different runs in the ImageCLEF medical image retrieval task using the test collections developed for the 2005 and 2006 tasks. The purpose of doing this work was to determine whether changes in relevance judgments significantly affect results, whether fusion of multiple runs can improve performance, and whether substituting frequency of retrieved images in runs can possibly substitute for human relevance judgments. We describe three sets of experiments with the ImageCLEF 2005 and 2006 medical test collections: (a) impact of varying levels of relevance and approaches to duplicate judgments, (b) impact of data fusion from multiple runs, and (c) impact of results derived from non-human judgments.

1 Background

One of the most difficult part of building test collections, and certainly the most resource-intensive aspect, is human relevance judgments. Not only do judgments cost money, but there is also concern over disagreement among judges and its impact on results. Voorhees has found, in the context of document retrieval in TREC, that different relevance judgments tend to give different absolute but comparable results [1]. Soboroff et al. have assessed whether randomly selected documents based on the distribution of known relevant documents in a collection could substitute for human judgments, finding that relative orders were maintained except at the high and low end of performance [2]. More recently, Aslam et al. developed a better sampling approach to be able to reproduce relative orders of results [3]. This paper describes variations on these experiments that we carried out using the test collections and submitted runs from the ImageCLEF 2005 and 2006 medical tasks.

The ImageCLEF medical image retrieval tasks for 2005 and 2006 were based on a library of about 50,000 images annotated in a variety of formats and languages and derived from four sources. The structure and annotation of the collection has been described elsewhere [4]. In 2005, there were 25 topics for the test collection consisting of a textual information needs

statement and an index image. The topics were classified posthoc into categories reflecting whether they were more amenable to retrieval by visual, textual, or mixed algorithms. Eleven topics were visually oriented (1-11), 11 topics were mixed (12-22), and three topics were semantically oriented (23-25). For 2006, more explicitly developed topics classified as amenable to retrieval by visual, textual, or mixed methods were developed. A total of 30 topics were developed, with 10 in each category.

In 2005, groups were required to classify runs based on whether the run used manual modification of:

- Queries input into systems - automatic vs. manual
- Retrieval methods - visual vs. textual vs. mixed

The two categories of topic modification and three categories of retrieval system type led to six possible run categories to which a run could belong (automatic-visual, automatic-textual, automated-mixed, manual-visual, manual-textual, and manual-mixed). For 2006, another category of topic modification was added, which was interactive. Manual modification meant that the query was modified from the topic by a human without looking at system output, whereas interactive modification meant that the query was modified based on viewing system output. This led to nine possible run categories (automatic-visual, automatic-textual, automated-mixed, manual-visual, manual-textual, manual-mixed, interactive-visual, interactive-textual, and interactive-mixed).

The final component of the test collections were the relevance judgments. As with most challenge evaluations, the collection was too large to judge every image for each topic. So as is commonly done in IR research, "pools" of images for judging each topic were developed, consisting of the top-ranking images in the runs submitted by participants [5].

Table 1 lists a variety of statistics from the 2005 and 2006 tracks, including the number of research groups, the number of runs they submitted, the top number of images used to construct the pools, the average pool size per topic, the total number of images judges, and the number of duplicates judged. The relevance assessments for both years were performed by physicians who were also graduate students in Oregon Health & Science University (OHSU) biomedical informatics program. All of the images for a given topic were assessed by a single judge using a three-point scale: definitely relevant, possibly relevant, and not relevant. The number of topics assessed by each judge varied depending on how much time they had available. Some judges also performed duplicate assessment of other topics.

Table 1 - Characteristics of data from 2005 and 2006

Attribute	**2005**	**2006**
Research groups	13	10
Runs submitted	134	100
Top images for pools	40	30
Average pool size per topic	892 (470-1167)	910 (647-1187)
Images judged	21,795	27,306
Duplicates judged	9,279	11,742
Runs analyzed	27	25

Once the relevance judgments were done, the results of the experimental runs submitted by participants were calculated using the trec_eval evaluation package (version 8.0, available from trec.nist.gov), which takes the output from runs (a ranked list of retrieved items for each topic) and a list of relevance judgments for each run (called qrels) to calculate a variety of relevance-based measures on a per-topic basis that are then averaged over all the topics in a run. The trec_eval package includes MAP (our primary evaluation measure), binary preference (B-Pref) [6], precision at the number of relevant images (R-Prec), and precision at various levels of output from 5 to 1000 images (e.g., precision at 5 images, 10 images, etc. up to at 1000 images).

2 Varying Relevance and Duplicate Judgments

One research question in this study asked whether the relative or absolute results of the submitted runs might be changed by varying the relevance judgments. This was done in two ways. One was to assess different levels of strictness for relevance. We assessed the impact on the results of runs for strict (definitely relevant only) versus lenient (definitely or possibly relevant) relevance.

Second, we looked at the impact of variation in relevance judgments. In both 2005 and 2006, about 40% of images were judged in duplicate. This not only allowed measurement of the consistency of the judging processed, but also provided us additional ways to alter the relevance judgments to assess the impact of variability. For both strict and lenient levels of relevance, we performed a Boolean AND of duplicated judgments (i.e., choosing the lowest level of relevance) and a Boolean OR (i.e., choosing the highest level of relevance). This provided in total six sets of qrels for trec_eval.

Table 2 shows the overlap of judgments between the original and duplicate judges. Judges were more often in agreement at the ends (not relevant, relevant) than the middle (partially relevant) of the scale. The kappa score, which measures chance-corrected agreement [7], was found to be in the range that statisticians define as “good” agreement.

In both years, a large number of runs were submitted for official scoring, many of which consisted of minor variations on the same technique, e.g., substitution of one term-weighting algorithm with another. We therefore limited our analysis of results to the best-performing run in a given run category from each group. This resulted in 27 runs analyzed in 2005 and 25 runs analyzed in 2006. Table 3 shows the run name, results, and type for the 27 analyzed runs from 2005, while Figure 1 shows the results plotted graphically and sorted by the “official” MAP, which in 2005 was based on strict relevance. Table 4 and Figure 2 show the same data for the 25 analyzed runs from 2006, although the “official” MAP for 2006 was calculated from lenient relevance.

Table 2 - Overlap of relevance judgments for (a) 2005 and (b) 2006.

(a) 2005 (Kappa = 0.679)

Duplicate	Relevant	Possibly relevant	Not relevant	Total
Original				
Relevant	1022	94	102	1218
Possibly relevant	157	83	153	393
Not relevant	236	199	7233	7668
Total	1415	376	7488	9279

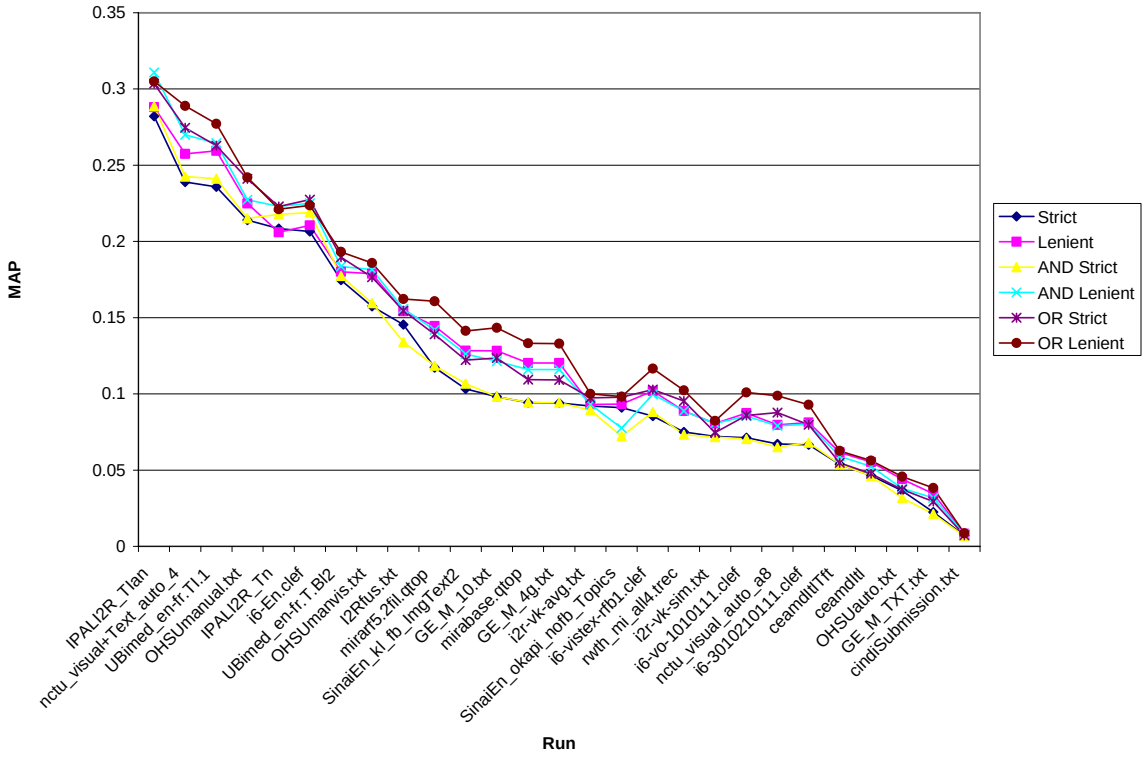
(b) 2006 (Kappa = 0.611)

Duplicate	Relevant	Possibly relevant	Not relevant	Total
Original				
Relevant	985	200	224	1409
Possibly relevant	282	91	433	806
Not relevant	171	186	9170	9527
Total	1438	477	9827	11742

Table 3 - Varying results for 2005 with strict and lenient qrels, combined by AND or OR with duplicated judgments.

Run	Type	Strict	Lenient	AND Strict	AND Lenient	OR Strict	OR Lenient
IPALI2R_TIan	AM	0.2821	0.2881	0.2887	0.3109	0.3034	0.3049
nctu_visual+Text_auto_4	AM	0.2389	0.2574	0.2425	0.2699	0.2745	0.2889
UBimed_en-fr.TI.1	AM	0.2358	0.2594	0.2412	0.2644	0.2628	0.2771
OHSUmanual.txt	MT	0.214	0.2249	0.2152	0.2273	0.2411	0.242
IPALI2R_Tn	AT	0.2084	0.206	0.2177	0.223	0.2229	0.2211
i6-En.clef	AT	0.2065	0.2106	0.2189	0.2250	0.2275	0.2237
UBimed_en-fr.T.Bl2	AT	0.1746	0.1801	0.1774	0.1836	0.1896	0.1931
OHSUmanvis.txt	MM	0.1574	0.1789	0.1595	0.1815	0.1766	0.1859
I2Rfus.txt	AV	0.1455	0.1542	0.134	0.1559	0.1545	0.1623
mirarf5.2fil.qtop	AM	0.1173	0.1446	0.1185	0.142	0.139	0.1607
SinaiEn_kl_fb_ImgText2	AM	0.1033	0.1283	0.1068	0.1262	0.1222	0.1413
GE_M_10.txt	AM	0.0981	0.1282	0.0981	0.1215	0.1235	0.1433
mirabase.qtop	AV	0.0942	0.1203	0.0943	0.116	0.1093	0.1332
GE_M_4g.txt	AV	0.0941	0.1202	0.0942	0.1159	0.1092	0.133
i2r-vk-avg.txt	MV	0.0921	0.0932	0.0894	0.0931	0.0974	0.1
SinaiEn_okapi_nofb_Topics	AT	0.091	0.0933	0.0722	0.0776	0.0978	0.0983
i6-vistex-rfb1.clef	MM	0.0855	0.1019	0.0881	0.0998	0.1028	0.1166
rwth_mi_all4.trec	AV	0.0751	0.0888	0.0733	0.0888	0.0953	0.1023
i2r-vk-sim.txt	AV	0.0721	0.0806	0.0717	0.0806	0.0746	0.0824
i6-vo-1010111.clef	AV	0.0713	0.0875	0.0705	0.0855	0.0859	0.101
nctu_visual_auto_a8	AV	0.0672	0.0797	0.065	0.0791	0.0877	0.0988
i6-3010210111.clef	AM	0.0667	0.0813	0.068	0.0802	0.0798	0.0929
ceamdItlTft	AM	0.0538	0.0617	0.0538	0.059	0.0548	0.0626
ceamdItl	AV	0.0465	0.0554	0.0462	0.0525	0.0476	0.0563
OHSUauto.txt	AT	0.0366	0.0442	0.0317	0.038	0.0373	0.0457
GE_M_TXT.txt	AT	0.0226	0.0346	0.0213	0.0318	0.0294	0.0384
cindiSubmission.txt	AV	0.0072	0.0084	0.0067	0.0081	0.0073	0.0087

Figure 1 - Graphical depiction of results from Table 3 for 2005.



3 Data Fusion

A second question this research addressed was whether data fusion (i.e., fusion of retrieved images) from multiple runs would produce results that might exceed the best single-system runs. This would not only determine whether the combination of the results from many runs might exceed the best single run, but also give us help in determining the optimal parameters for the final set of experiments described below.

In order to assess this question, we built retrieval sets by taking the top N images retrieved from each analyzed run and sorted them by the frequency in those top N. Table 5 shows the official MAP for each year for N at various levels, while Figure 3 depicts the results graphically. For both years, the fused runs exceed the best MAP for any single run (0.2881 in 2005 and 0.3095 in 2006). The performance with very small N is poor, but quickly rises to a peak and then tapers off or slightly falls. The peak is reached sooner for 2005 than 2006 data but nonetheless for both exceeds the single best run.

Table 4 - Varying results for 2006 with strict and lenient qrels, combined by AND or OR with duplicated judgments.

Run	Type	Lenient	Strict	AND Lenient	AND Strict	OR Lenient	OR Strict
IPAL-IPAL_Cpt_Im	AM	0.3095	0.2959	0.2974	0.2833	0.3068	0.3123
IPAL-IPAL_Textual_CDW	AT	0.2646	0.2488	0.2483	0.2358	0.2754	0.259
IPAL-IPAL_Textual_CRF	FT	0.2534	0.2345	0.2375	0.2169	0.2533	0.2421
GE_8EN.treceval	AT	0.2255	0.2252	0.2121	0.2139	0.2374	0.2303
OHSUeng	MT	0.2132	0.1983	0.2029	0.199	0.2192	0.2049
UB-UBmedVT2	AM	0.2027	0.1905	0.1947	0.1875	0.2107	0.1935
UB-UBmedT1	AT	0.1965	0.1907	0.1778	0.179	0.206	0.1943
UKLFR-UKLFR_origmids_en_en	AT	0.1698	0.1595	0.1512	0.1515	0.1731	0.1642
RWTHi6-EnFrGePatches	AM	0.1696	0.1467	0.1572	0.1415	0.1728	0.153
IPAL-IPAL_CMP_D1D2D4D5D6	MV	0.1596	0.151	0.1501	0.1467	0.1673	0.1609
OHSU-OHSU_m1	FM	0.1563	0.1216	0.1492	0.1233	0.1548	0.1305
RWTHi6-En	AT	0.1543	0.1347	0.1384	0.125	0.1583	0.1399
cindi-CINDI_Text_Visual_RF	FM	0.1513	0.159	0.1434	0.1511	0.1508	0.1585
OHSU_baseline_trans	AT	0.1264	0.1077	0.1203	0.0943	0.1274	0.1095
GE_vt10.treceval	AM	0.12	0.1128	0.1173	0.1163	0.1269	0.1156
SINAI-SinaiOnlytL30	AT	0.1178	0.1003	0.1091	0.0911	0.1213	0.1045
cindi-CINDI_Visual_RF	FV	0.0957	0.0871	0.0958	0.0868	0.0966	0.0945
cindi-CINDI_Fusion_Visual	AV	0.0753	0.0763	0.0751	0.0754	0.0772	0.0796
MSRA_WSM-msra_wsm	AV	0.0681	0.0711	0.0703	0.0697	0.0741	0.0741
IPAL-IPAL_Visual_SPC+MC	AV	0.0634	0.0588	0.0619	0.0547	0.0716	0.0641
INSA-CISMef	MM	0.0531	0.05	0.0557	0.0496	0.053	0.0517
RWTHi6-SimpleUni	AV	0.0499	0.045	0.0497	0.0427	0.0555	0.0467
GE-GE_gift	AV	0.0467	0.0431	0.0462	0.042	0.0546	0.0459
SINAI-SinaiGiftT50L20	AM	0.0467	0.0431	0.0462	0.042	0.0546	0.0459
UKLFR-UKLFR_mids_en_all_co	AM	0.0167	0.0139	0.0151	0.013	0.0177	0.0151

Figure 2 - Graphical depiction of results from Table 4 for 2006.

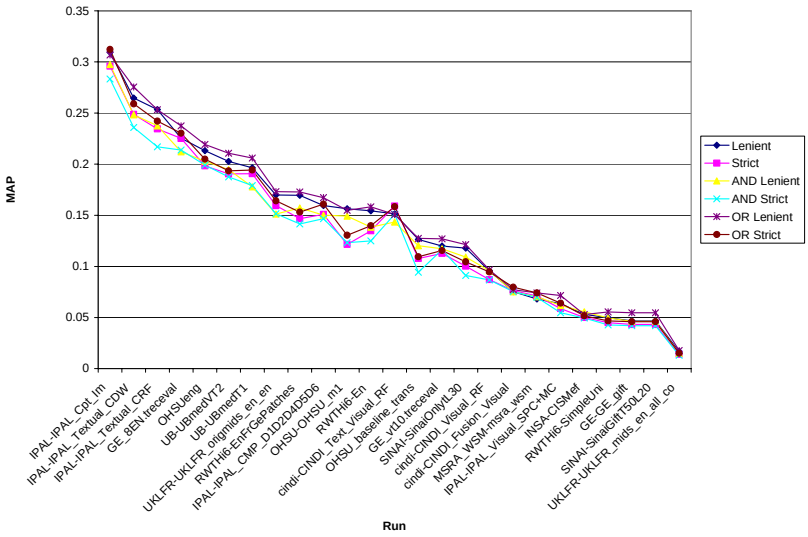
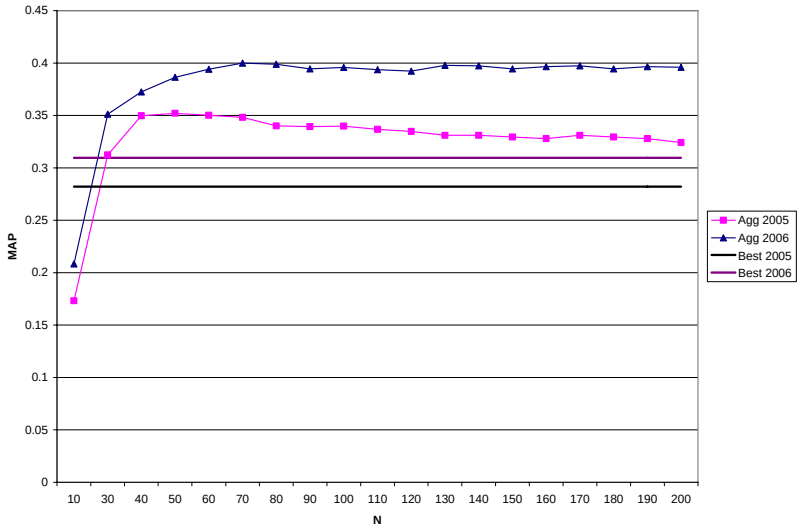


Table 5 - Data fusion results for 2005 and 2006 by varying levels of frequency of image occurring in top N results of each run.

N	2005	2006
10	0.1732	0.2083
30	0.3122	0.3513
40	0.3498	0.3724
50	0.3521	0.3864
60	0.3501	0.3941
70	0.3482	0.4
80	0.34	0.3988
90	0.3393	0.3944
100	0.3397	0.3959
110	0.3368	0.3937
120	0.3347	0.3923
130	0.331	0.3977
140	0.331	0.3973
150	0.3293	0.3944
160	0.3279	0.3965
170	0.331	0.3973
180	0.3293	0.3944
190	0.3279	0.3965
200	0.3241	0.396
300	0.3029	0.3922
500	0.2816	0.3727
Best	0.2821	0.3095

Figure 3 - Graphical plot of Table 5 for 2005 and 2006 also showing best individual run for those years.



4 Results based on non-human relevance judgments

The final question addressed by this research looked at the impact of generating qrels based on the frequency of images retrieved. The major motivation for this approach is whether it could replace the need for costly human relevance judgments. We assessed this question by building new qrels files by varying two parameters: the number of runs where images appeared in the top N and the size of the qrels files. The potential variation for both of these parameters is limitless, so we used the data fusion experiments to guide us.

Figures 4 and 5 show the results of substituting these qrels for human judgments. We varied whether qrels were gathered from the top 10, 30, 50, or 100 ranked images from the official runs as well as the size of qrels file from the highest ranking 30, 60, or 100 images. The results obviously show that this simple approach is a poor substitute for human judgments. While the automated results tend to be consistent among themselves, they do not predict the absolute or relative performance well at all.

Figure 4 - Results from substituted qrels based on frequency of image occurrence for 2005.

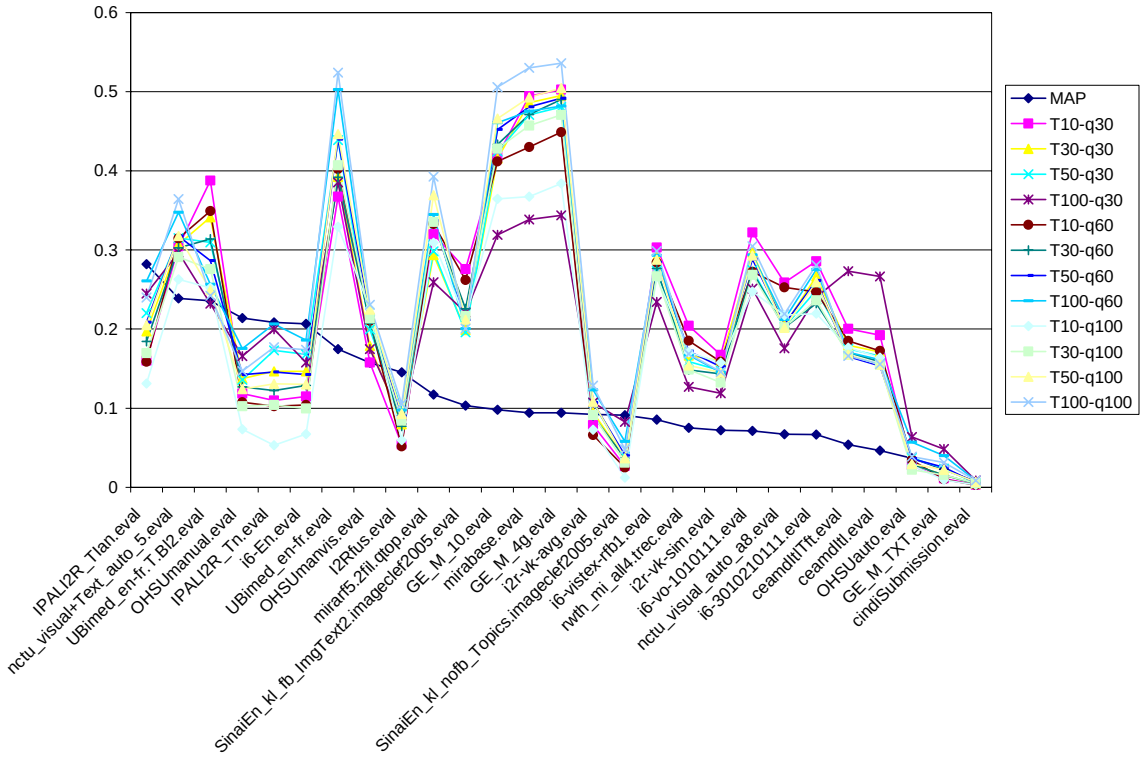
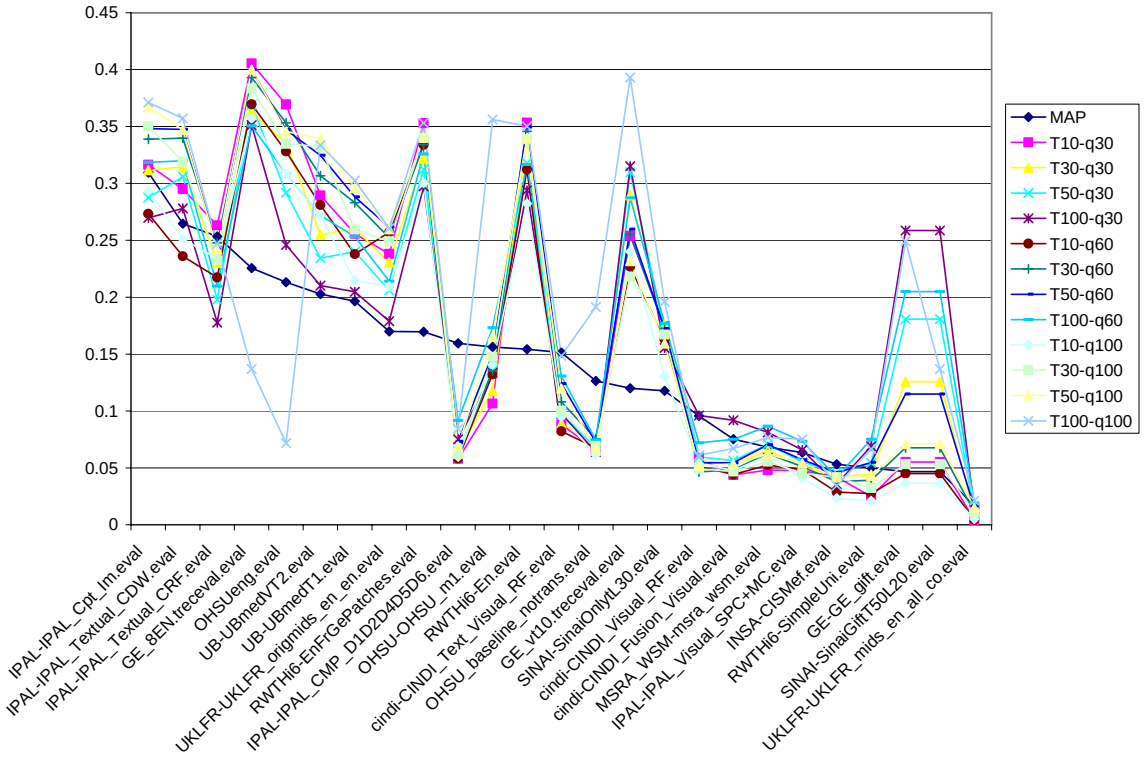


Figure 5 - Results from substituted qrels based on frequency of image occurrence for 2006.



5 Conclusions

In this paper, we explored the impact of varying relevance judgments on the results from the ImageCLEF medical image retrieval task. We found that varying relevance judgments based on lenient vs. strict relevance or via different interpretations of duplicate judgments had small absolute and relative impacts on results, indicating that results are relatively robust to varying relevance judgments. We also discovered that data fusion from different runs leads to better overall performance than any individual run, indicating potential promise for systems that incorporate output from multiple systems or algorithms. Finally, we found that simply replacing human-generated qrels with those based on frequency of retrieved images were not an effective replacement for real judgments. Clearly other approaches are needed if human judgments are to be replaced.

Acknowledgements

This work was funded in part by Grant ITR-0325160 of the US National Science Foundation.

References

1. Voorhees EM. Variations in relevance judgments and the measurement of retrieval effectiveness. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1998. Melbourne, Australia: ACM Press. 315-323.
2. Soboroff I, Nicholas C, and Cahan P. Ranking retrieval systems without relevance judgments. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2001. New Orleans, LA: ACM Press. 66-73.
3. Aslam JA, Pavlu V, and Yilmaz E. A statistical method for system evaluation using incomplete judgments. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2006. Seattle, WA: ACM Press. 541-548.
4. Hersh WR, et al., Advancing biomedical image retrieval: development and analysis of a test collection. *Journal of the American Medical Informatics Association*, 2006: Epub ahead of print. http://www.jamia.org/cgi/reprint/M2082v1.
5. Buckley C and Voorhees EM, Retrieval System Evaluation, in *TREC: Experiment and Evaluation in Information Retrieval*, Voorhees EM and Harman DK, Editors. 2005, MIT Press: Cambridge, MA. 53-75.
6. Buckley C and Voorhees EM. Retrieval evaluation with incomplete information. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2004. Sheffield, England: ACM Press. 25-32.
7. Cohen J, A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960. 20: 37-46.