A Comparative Analysis of Human and Automatic Query Variants

Binsheng Liu
RMIT University
Melbourne, Australia

Nick Craswell
Microsoft
Redmond, USA

Xiaolu Lu
RMIT University
Melbourne, Australia

Oren Kurland
Technion – Israel Institute of Technology
Haifa, Israel

J. Shane Culpepper
RMIT University
Melbourne, Australia

ABSTRACT

We present an in-depth comparative analysis of the effectiveness distributions of sets of human-created and automatically-created query variations used to represent the same information need. The automatic variations are generated using Bing's click graph. Experiments performed with TREC datasets show that using automatic variations for retrieval can result in similar effectiveness to that of using human variations, although the two types of variations can be appreciably different in several important respects — e.g., their similarities and corresponding retrieved lists.

ACM Reference Format:
Binsheng Liu, Nick Craswell, Xiaolu Lu, Oren Kurland, and J. Shane Culpepper. 2019. A Comparative Analysis of Human and Automatic Query Variants. In *Proceedings of The 5th ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '19).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3341981.3344223

1 INTRODUCTION

The distinction between a query and the underlying information need represented by the query has been an essential component of Information Retrieval research for more than half a century. Many factors influence the effectiveness of a keyword query, and small reformulations can have a substantial impact on the performance. While several studies have explored the effectiveness gap between human query formulations and automatically generated variations [1, 7], recent work has renewed interest in this fundamental IR problem [2–5]. However, manually curated collections of queries do not necessarily translate to performance improvements in a production setting where related queries discovery must somehow be operationalized.

In this work, we comprehensively compare *sets* of automatically generated query variants — produced using random walks over click graphs derived from Bing query logs — and human generated variants using two commonly used TREC test collections. This differs from previous studies which made comparisons on a query-by-query basis instead of comparing performance distributions of multiple queries for a single information need simultaneously. Wefi nd that the retrieval effectiveness of the automatic variants can reach the level of human-created variants, although the two sets of variants (per topic) are quite different in several respects: the variants themselves, their similarities and the corresponding retrieved lists. So, the gap still exists, promising advances in IR and closely relatedfields such as natural language processing and machine translation may finally make automatic query variant generation a reality.

Ourfi ndings have important implications: human-created variants can serve as surrogates for automatically-generated variants and vice versa in terms of retrieval effectiveness. At the same time, their differences motivate further research along the direction of creating and understanding different forms of information need representations.

Research Questions. In this paper, we perform a comparative analysis of human generated and automatically-generated query variations, and in particular, our experiments explore the following research questions: (1) Can automatically generated query variations be as effective for retrieval as carefully crafted human query formulations? (2) What are the similarities and differences between the variations being produced?

2 PRELIMINARIES

Document Collections and Retrieval Models. We use two test collections in our experiments: ROBUST (disk 4&5 - CR), and the 2013–2014 Web Track ClueWeb12 CatB (CW12B) document collections. The Indri toolkit is used for all retrieval experiments, and stopwords were pruned from queries at runtime. Across all experiments, Krovetz stemming is applied to queries and documents, and query likelihood model (QL) [9] (Dirichlet smoothed document language model, $\mu = 2500$) was used for retrieval.

Query Variations. Instead of using title queries as originally provided by TREC, for each topic, we consider two sets of query variants. Thefi rst set were manually curated, human query variations created through crowdsourcing experiments. The second set were generated automatically using a random walk on a click graph derived from query logs in the Bing search engine. Human generated query variations for both CW12B and ROBUST are publicly available, and have been used in several recent research papers. [1,2]

The automatic query variants are generated by using a bipartite query–URL click graph taken from a 10% sample of Bing click data over several months in 2018. Note that the automatic variants are queries selected from the log as those presumably most related to the query at hand. Sheldon et al. [8] proposed a process to induce multiple query variations from a starting query or description using the random walk model originally described by Craswell and Szummer [6]. Using a two step forward walk produces queries that would be reached if the walk starts with a single user query. Here we apply the same model, but use a two step backward walk, which tells us what queries were the likely starting point given that we ended at the user query. The backwards walk model also performed better in the

ICTIR '19, October 2–5, 2019, Santa Clara, CA, USA

ACM ISBN 978-1-4503-6881-0/19/10…$15.00
https://doi.org/10.1145/3341981.3344223

[1] https://culpepper.io/publications/robust-uqv.txt.gz
[2] http://dx.doi.org/10.4225/49/5726E597B8376

Table 1: Retrieval effectiveness; for Bing, median performance is reported for different filtering thresholds of bottom-performing queries. All statistical significance tests are performed against median queries in both query sets. † and ‡ mean $p < 0.05$ in the t-test and TOST test ($\Delta\text{AP} = 0.05$) compared to title query, respectively. *h* and *b* mean $p < 0.05$ in the t-test compared to human best and bing best respectively.

Query Set		CW12B			ROBUST		
		MAP	NDCG@10	RBP@0.95	MAP	NDCG@10	RBP@0.95
Title query	-	0.201	0.192	0.360+0.213	0.247	0.426	0.308+0.035
Human	Median	0.178	0.190	0.351+0.185	0.239	0.421	0.294+0.124
Bing Median for threshold	0.0	$0.103^{\dagger}$	$0.120^{\dagger}$	$0.230+0.495^{\dagger}$	$0.144^{\dagger}$	$0.254^{\dagger}$	$0.179+0.364^{\dagger}$
	0.1	$0.118^{\dagger}$	$0.138^{\dagger}$	$0.247+0.445^{\dagger}$	$0.160^{\dagger}$	$0.296^{\dagger}$	$0.204+0.325^{\dagger}$
	0.3	$0.141^{\dagger}$	$0.146^{\dagger}$	$0.284+0.395^{\dagger}$	$0.182^{\dagger}$	$0.337^{\dagger}$	$0.226+0.289^{\dagger}$
	0.5	$0.166^{\ddagger}$	$0.192^{\ddagger}$	$0.323+0.313^{\dagger}$	$0.201^{\dagger}$	$0.358^{\dagger}$	$0.248+0.249^{\dagger}$
	0.7	$0.194^{\dagger}$	0.210	0.366+0.271	$0.228^{\ddagger}$	0.402	$0.281+0.216^{\ddagger}$
	0.9	$0.226^{\dagger}$	$0.243^{\dagger}$	$0.407+0.218^{\dagger}$	$0.273^{\dagger}$	$0.466^{\dagger}$	$0.330+0.174^{\dagger}$
Human	Best	0.286	0.304	0.501+0.118	0.373	0.604	0.422+0.078
Bing	Best	0.239	0.252	0.428+0.215	0.282	0.481	0.338+0.170
Combined	Best	0.288^{b}	0.303^{b}	$0.503+0.120^{b}$	$0.389^{h,b}$	$0.621^{h,b}$	$0.436+0.081^{h,b}$

original paper [6], and produced the best results in our preliminary tests. We did not perform additional experiments to pick the best hyper-parameters for the random walk, and leave this for future work. We note that for description queries, the query is very unlikely to occur in the graph, so temporary nodes were created for each description query that was connected to any URLs found in the description query's top-50 Bing results. Note that the descriptions used for CW12B were the backstories developed by Bailey et al. [2], and TREC descriptions were for ROBUST to ensure that the queries being generated were directly comparable to the human-generated query sets.

As a result, on average, there are around 16 automatic query variations and 12 human-generated query variations for ROBUST; for CW12B, there are around 25 automatic and 39 manual query variations available. There are in total 100 topics for CW12B, and 249 topics for ROBUST. For automatic variants, none were produced for three topics in ROBUST, and so these were treated as empty queries in all comparisons to ensure that our results are directly comparable to previous results reported for ROBUST.

Experimental Methodology. We now describe our experimental methodology. First, QL retrieval is performed using each query variation over the corresponding document collections, and then AP (average precision) is computed using trec_eval. Every query variation with a 0 AP was dropped (on average 2 queries were dropped per topic from the human set using this methodology, and 4 from the automatic set). This is consistent with previous work on UQVs [3, 4]. The goal of our work is to better understand how variations of similar quality that were manually generated by humans compare to automatically generated ones.

To address our first question, we gradually remove the bottom-performing x% of the Bing queries for a topic, in order to find a set of automaticly generated (selected) queries that are of comparable effectiveness quality to the human curated set. The equivalence between the two sets of queries is determined based on the median queries from the human reference sets, and the median of the current set of pruned automatic queries. A paired, two-sided t-test and a two one-sided test (TOST) can be used to identify the most appropriate cutoff threshold. TOST is commonly used in the medical community to test for statistical non-inferiority [10]. More specifically, a t-test tests for differences while a TOST tests for equivalence. Once a cutoff is identified in the automatic collection which results in a similar effectiveness to the human collection, the two query sets are then exhaustively compared and contrasted.

In order to address the second research question, we explore the similarity between queries for a topic (Intra) and also compare similarity between the queries for Human and Bing (Inter). We use Jaccard similarity and *Rank-Biased Overlap* (RBO) [11] to do our intra- and inter- similarity comparisons, which is discussed more in the next section.

3 RESULTS AND FINDINGS

Retrieval Performance. We show results for the automatic variations at different filtering cutoffs in Table 1. On average, both tests suggest that the (median) AP effectiveness of automatic and human variants is statistically indistinguishable after pruning the bottom 50% queries in CW12B. We see a similar trend for ROBUST when pruning the bottom 70%. This observation is not surprising: Table 1 shows that the human variations for ROBUST are of higher quality than those for CW12B. Human variants for ROBUST were created by search domain experts, while those for CW12B were created through an online crowdsourcing experiment.

The most important take-away message from this comparison is that small perturbations of a query can have a significant impact on performance. For example, the original TREC title query for Bing-ROBUST is the best variant in terms of MAP for 73 of 249 queries (29%), and for Bing-CW12B, 14 of 100 queries. For human variants, it is 23 of 249, and 0 of 100 respectively. So, even in our first attempt, the original TREC title query is superior for only 1/3 of the topics. Furthermore, fusion of variants consistently outperforms a single query, even when very few variants are available [3, 4]. We do not explore this further here due to space limitations.

To answer our first question, our results show that automatic queries may be able to achieve a similar performance level to human

Table 2: Query Jaccard Similarity: within a query set (Intra), between the query sets (Inter) and with TREC's topic Title, B and H stand for Bing and Human, respectively.

	Set	Intra Sim.		Inter Sim.		Sim. to Title	
		Avg	Max	Avg	Max	Avg	Max
CW12B	B	0.372	0.917	0.299	0.971	0.436	0.825
	H	0.331	0.820			0.407	0.916
ROBUST	B	0.299	0.699	0.190	0.600	0.286	0.524
	H	0.312	0.730			0.335	0.701

generated ones, but a gap still exists in the percentage of low quality variants being induced through our automatic generation approach. Moreover, we can not ignore the fact that the residuals of RBP@0.95 are much larger than anticipated on both sets of queries, indicating the high level of uncertainties in our comparison, and lower than expected judgment coverage in both collections.

Topic Differences. Retrieval effectiveness varies on a per topic basis, as do the query variations themselves, and these differences can be observed directly in Figure 1, where we plot performance of pruned automatic variations and human variations relative to the median query of all known variants for that collection. We also show where the original TREC title query performance lies for that topic. As we can see, variant performance on both sets varies widely for nearly every topic. Since Figure 1 suggests that there are large quality differences between query variations at the topic level, the relative differences can be further quantified using the *drop rate* shown in Figure 2, which is the percentage of query variations that must be pruned per topic from the automatic set to make the two sets statistically indistinguishable.

We can see that, although on average, CW12B drops 53.8% automatic queries and ROBUST drops 61.9% to be similar to human variations, this percentage actually varies significantly per topic: 13 topics on CW12B and 30 topics on ROBUST in the automatic set outperform the human reference set, while 12 topics on CW12B and 74 topics on ROBUST cannot achieve similar performance to the human benchmark. This is an interesting observation for several reasons. First it suggests that both approaches can produce effective queries – in fact queries that are more effective for the information need than previously known, i.e., TREC topic titles. It also suggests that despite similar overall performance, the two methods are capable of producing remarkably different query variations. This motivates us to further explore the diversity exhibited by the two sets.

Term-Level Similarity. We begin to answer our second research question by exploring the similarity at the term-level. To accomplish this, we first want to determine how many queries were exactly duplicated in the human versus automatic query sets. On average, 2.7 queries match in CW12B and 0.2 in ROBUST, with at most 8 and 2 matches for any one topic respectively in the two collections.

Next we extend our comparison intra-similarity: the similarity within each set of query variations. We measure the Jaccard similarity between the terms in the queries to further quantify the overlap. As shown in Table 2, automatic query variations on CW12B exhibit a slightly higher similarity than human variations on average per topic — with similarities as high as 0.917 for some topics. However, ROBUST behaves differently. Here, the human generated query variations have a higher similarity. In addition, we also found that both sets of variations tend to be quite similar to TREC title queries. This is perhaps not terribly surprising as TREC title queries were formulated by topic originators, and the descriptions are simply an exposition of that intent. We also observed that automatic variations have a marginally higher similarity to title queries on CW12B than for ROBUST, which we hope to explore this further in future work.

Table 3: Retrieval consistency measured using RBO.

	Set	Intra Sim.		Inter Sim.		Sim. to Title	
		Avg	Max	Avg	Max	Avg	Max
CW12B	B	0.270	1.000	0.158	1.000	0.346	1.000
	H	0.162	1.000			0.223	1.000
ROBUST	B	0.382	1.000	0.205	1.000	0.329	1.000
	H	0.233	1.000			0.323	1.000

Finally, we consider inter-similarity: the similarity between the automatic and human generated queries. The average inter-similarity scores shown in Table 2 suggest that the two sets of queries also express the same information need in different ways, with an average similarity of 0.299 and 0.190 on CW12B and ROBUST, respectively. Again, there is a difference between the two test collections: the set of human queries and automatic queries are more similar on CW12B than on ROBUST, given that the maximum similarity can reach 0.971 on CW12B, but only 0.600 on ROBUST. It is worth noting that the inter-similarity is lower than intra-similarity on both collections, which reinforces the inherent differences of the two sets in expressing the same information need.

Comparing Retrieval Similarities. In order to gain a better understanding of the differences between the two sets of queries, we turn now to study retrieval consistency measured with RBO, $p = 0.9$ [11], as shown in Table 3, which measures similarities between two ranked retrieval lists. In general, the retrieval similarity is correlated with the term similarity in Table 2, with the exception of ROBUST, where automatic queries have a slightly lower level of query-level similarity but a slightly higher RBO score. There could be many reasons for this discrepancy, for example, the query length, which differs in the two collections. When considering the inter-set similarity, we observe a very low agreement between the two sets, implying that documents retrieved using the two sets of variants are in general highly diverse.

Example Queries. For some topics, human variants and Bing variants are superior in capturing different aspects of the information need, often complementing each other. We now perform a qualitative analysis for a few interesting topics on where the two sets of variants behave very differently. We found that human variants are better at addressing information needs that involve rare words or common misspellings. For ROBUST topic 301 "agoraphobia" and topic 677 "leaning tower pisa", human variants outperform Bing variants, as most human variants contain the correct word "agoraphobia" and "pisa", while Bing variants rarely contain the title query or contain misspelling such as "pizza". The difference stems from the generation methods for the two sets. Human variants were collected through crowdsourcing (or domain experts), during which workers were able to see the exact keyword in the information need description and then provide correctly spelled queries. Conversely, Bing variants can show a deeper understanding and capture domain knowledge as they are induced from real user queries. For example, ROBUST Topic 372 is "native american casino". Human variants were "native

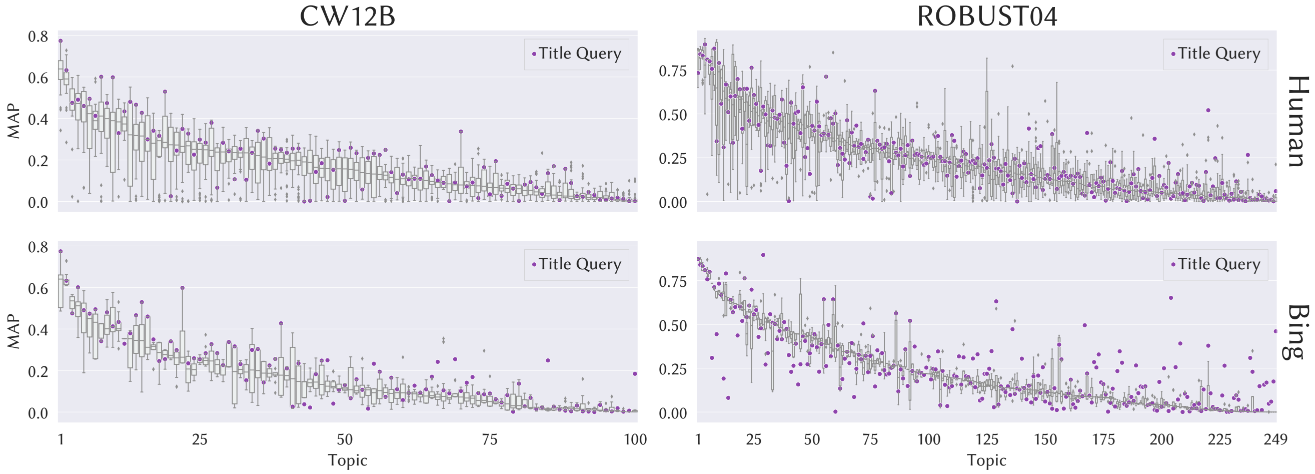

Figure 1: Per-topic comparisons. Automatic query variations are in the pruned set, where the pruning percentages are 50% and 70% on CW12B and ROBUST, respectively.

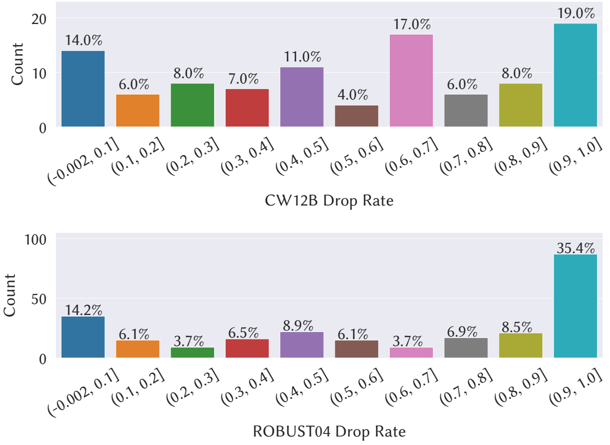

Figure 2: Per-topic drop rate of automatic query variations. The x-axis is the drop rate and the y-axis is the number of dropped variants.

american casino gambling", "tribe casino gambling", and "indigenous peoples america casino gambling" which are just synonyms of the title query, while Bing variants included "indian casino", "500 nations", and "igra" which have very high specificity. Bing variants also tend to be more natural queries than human variants as crowdsourcing workers can be unconsciously influenced by the description when choosing query terms. Topic 658 is "teenage pregnancy" which also appears in many human variants, but Bing variants usually contain "teen pregnancy" and yielded significantly higher AP scores. So it would appear that both automatic and human-based approaches can be used to produce query variations effectively, but can also result in queries with very different properties. Both approaches complement each other in unexpected ways, which becomes even more apparent when considering the "Best" human, Bing, and Combined effectiveness results shown at the bottom of Table 1.

4 CONCLUSIONS AND FUTURE WORK

We compared and contrasted two approaches to creating query variations for a single information need on two different commonly used test collections – manually by humans and automatically using data produced in a commercial search engine. We showed that while both human created and automatically generated variants can achieve comparable performance, subtle differences between the queries being created still exist. An important take-away message from our empirical analysis is that remarkable effectiveness gains are still possible based purely on the query formulation of an information need, in the automatic and human settings. Note that our preliminary experiments only explored performance using a single, simple bag-of-words ranking algorithm: query likelihood. Additional performance improvements can be achieved using more effective ranking algorithms (e.g., learning-to-rank). Understanding which query reformulations are effective for which retrieval approaches is an interesting research question that we plan to continue exploring in future work.

Acknowledgements. We thank the reviewers for their helpful comments. This work was supported in part by the Israel Science Foundation (grant no. 1136/17), the Australian Research Council's *Discovery Projects* Scheme (DP170102231), an Amazon Research Award, and a Google Research Award.

REFERENCES

[1] L. Azzopardi, M. de Rijke, and K. Balog. 2007. Building simulated queries for known-item topics: an analysis using six european languages. In *Proc. SIGIR*. 455–462.
[2] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. 2016. UQV100: A Test Collection with Query Variability. In *Proc. SIGIR*. 725–728.
[3] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. 2017. Retrieval Consistency in the Presence of Query Variations. In *Proc. SIGIR*. 395–404.
[4] R. Benham and J. S. Culpepper. 2017. Risk-Reward Trade-offs in Rank Fusion. In *Proc. ADCS*. 1–8.
[5] R. Benham, J. S. Culpepper, L. Gallagher, X. Lu, and J. Mackenzie. 2018. Towards efficient and effective query variant generation. In *Proc. DESIRES*. 62–67.
[6] N. Craswell and M. Szummer. 2007. Random walks on the click graph. In *Proc. SIGIR*. 239–246.
[7] R. Cummins, M. Lalmas, and C. O'Riordan. 2011. The limits of retrieval effectiveness. In *Proc. ECIR*. 277–282.
[8] D. Sheldon, M. Shokouhi, M. Szummer, and N. Craswell. 2011. LambdaMerge: Merging the Results of Query Reformulations. In *Proc. WSDM*. 795–804.
[9] F. Song and W. B. Croft. 1999. A general language model for information retrieval (poster abstract). In *Proc. of SIGIR*. 279–280.
[10] E. Walker and A. S. Nowacki. 2011. Understanding equivalence and noninferiority testing. *J. General Internal Medicine* 26, 2 (2011), 192–196.
[11] W. Webber, A. Moffat, and J. Zobel. 2010. A Similarity Measure for Indefinite Rankings. *ACM Trans. Information Systems* 28, 4 (Nov. 2010), 20.1–20.38.