Task 2a: Team KU-CS: Query Coherence Analysis for PRF and Genomics Expansion

Ornuma Thesprasith and Chuleerat Jaruskulchai

Department of Computer Science, Faculty of Science,
Kasetsart University, Thailand
ornuma.thesprasith@gmail.com and fscichj@ku.ac.th

Abstract. Laypeople who are not medical expert may formulate short query using words from their discharge summaries or long query that explain their health conditions. The different query styles should be treated with different query expansion mechanisms. This work is an adaptive query expansion based on the coherence among query terms. To provide users with more readability documents, the document complexity is analiyzed on the word lenght and is used as a re-ranking method. The baseline retrieves using Lucene v.4.6 with default configuration. Overall retrieval performance of the baseline is better than the adaptive query expansion (P@10, MAP, and NDCG). Since the complexity of document uses only length of word, the readability performance after re-ranking is poorly performing.

Keywords: Adaptive query expansion ·Query coherence analysis · Genomics expansion ·Readability

1 Introduction

Health information is now available on the web and easy to access by non-professionals who associate to health care system and called laypeople. The 2015 CLEF eHealth [1] aims to help laypeople in seeking health information. To foster research and development of health-related search engine, the 2015 CLEF eHealth Task 2 [2] provides a shared collection of health-related web pages and queries set. This task assumes that laypeople formulate query with more terms to represent a single medical term. For example, "white part of eye turned green" means to medical term "jaundice" [3].

From the example, we assume that the medical term(s) should be presented in the relevant web page(s) and may be absented from the query. There is a gap between query and relevant web pages. Query expansion (QE) technique is commonly used to handle this situation. Since there are many approaches for doing expansion, we examine the effectiveness of different query expansion approaches for different query styles. This work examines three local-based and one global-based query expansion mechanisms. The local-based QE uses documents from the initial query retrieval and the global-based QE uses all documents from the collection for term selection.

The degree of coherence among query terms is used to estimate performance of the query retrieval, called QPPpair. We assume that the query with performing well should be expanded with terms from the small-size of top-ranked documents or no need to expand any term. Because these documents are assumed to be the relevance. The query with performing poorly should be expanded with terms from the entire collection or from an external source.

In addtion to retrieve more useful web pages, the results should be easy to read for laypeople. According to the readability requirement, factors used to evaluate the readability [4–6] are sentence length, number of sentences, the number of syllables or number of characters per word. Since our work treats the collection as a bag of words therefore we assume that long word is a complex word and then the complexity of document can be estimated by the frequency of the long words.

In this paper, we use the following techniques to achieve the CLEF 2015 eHealth Task 2 [2] task; query performance prediction, query expansion and document readability. The existing and related works of each technique are briefly introduced in the next section. Our purposed method is described in the Section 3. The experiment setting is explained in the Section 4. The results and discussion are in the Section 5 and 6, respectively.

2 Related Works

2.1 Query Performance Prediction

Query performance prediction (QPP) [10] estimates retrieval effectiveness without relevance judgments. The utilization of the QPP method is not limit to estimate performance of query only, but also to determine query expansion mechanism and to select the most effectiveness expansion source as reported in works [7, 8].

The survey of pre-retireval QPP predictors [9] have organized the QPP predictors as specificity, ambiguity, term relatedness, and ranking sensitivity. The specificity of query is estimated from frequency of query term(s) in the collection and number of documents that term presents. For example, the Average Inverse Collection Term Frequency (AvICTF)[10] prefers terms that infrequenty appear in the collection. The query with high AvICTF value should be related to the documents that contain these specific terms and these documents may be the relevant documents. The work [8] has used the AvICTF [10] to adapt query expansion mechanisms; non-expansion, or expand with terms from collection that have the highest AvICTF value. Our work is inspired by this work [8] but we use different QPP measure.

We focus on the term relatedness of QPP predictor because we assume that query from laypeople contains more terms, as seen from the query example, therefore relation among query terms may be reflect to the query performance. The examples of the term relatedness-based QPP predictors [9] are the Average Pointwise Mutual Information (AvPMI), the Maximum Pointwise Mutual Information (MaxPMI) and the query coherence score[11].

2.2 Query Expansion Approaches

Query expansion (QE) is a technique which new terms are added to an original query and assumes that the new term(s) should be enlarging the document range for second-pass retrieval. There are two major factors that influence on the effectiveness of QE technique; term selection method and reweighting method[12]. Our work focuses on the first factor only by investigating different sources of terms for expansion. The term selection method based on local context analysis method[13] is reported in[14]. They constructed concept hierarchies from text corpus and then provided related concept(s) for users to select in interactive expansion manner. This idea is also applied in the construction of semantic network for interactive query expansion[15].

There are three steps of concept selection described in[15]. The first step is filtering concepts from terms. A concept is a term that frequently appears in the retrieved documents set respect to entire collection. The second step is finding the important concepts based on entropy of these concepts. The important concepts should not frequently appear or infrequently appear in the retrieved set. The entropy will be decreased if a concept is appeared more often or rarely. The third step is finding the related concepts based on conditional probability. Two concepts will be related if their conditional probabilities are more than a threshold.

2.3 Readability

Readability of health information is a current issue as seen in many current works examined the readability of web pages. Each work focuses on different health condition such as epilepsy[4], breast cancer[5], stroke[6]. The commonly used readability measures are the Fleshch-Kincaid grade level and the Simple Measure of Gobbledygook (SMOG). The factors commonly used in the readability formulas are sentence length, number of sentences, the number of syllables or characters per word. Although these studies are focused on different topics, their conclusion are in common. They found that health-related web pages have a high grade level of the readability which uneasy to understand by laypeople.

The alternative way to measure readability is using language model[16] . The model estimates the reading difficulty of web pages as the classification task. The reading level classifier is based on linear combination of language model and surface linguistic features in the document.

3 Methods

3.1 QPP Based on Query Term Coherence

Given an assumption that characteristic of the query of CLEF 2015 task[2] consists of more general words, we aim to measure the relatedness among query terms. If a pair of two terms occurs more often together in the corpus, then the pair is seem to be related. Since we have observed that non-stop words in one

line of scientific papers are approximately appeared 7-8 words. Therefore in this work we fixed window size at 7 word lenght for determining the pair.

Our QPP method is called the $QPPpair$, which measures how many pairs of query terms existing in the collection (called $PairExist$) respect to all possible pairs in the query (called $AllPairs$). The formula for measuring query coherence is defined as Equation 1.

$$QPPpair(Q) = \frac{\Sigma_{p \in Q} PairExist(p)}{AllPair(Q)} \tag{1}$$

$$PairExist(p) = \begin{cases} 1 & \text{, if pair exist in collection} \\ 0 & \text{, otherwise} \end{cases} \tag{2}$$

$$AllPair(Q) = \Sigma_{p \in Q} Pair(p) \tag{3}$$

In the preliminary experimental, we have examined the $QPPpair$ of query set in three collections; CLEF eHealth 2014, TREC Genomics 2004, and OHSUMED. The results show that the Genomics collection is more consistent with the CLEF collection so that we use the Genomics collection as external source of query expansion. We use the average of the $QPPpair$ value of three collections to determine query expansion mechanism.

3.2 Query Expansion Mechanisms

There are five query expansion mechanisms; no query expansion, expanded with top-small PRF set, expanded with top-medium PRF of both local and external collection, expanded with top-large PRF of both local and external collection, and expanded with co-occurrence terms in global collection.

Pseudo Relevance Feedback Query Expansion (PRF-QE). The retrieved documents contain at least one query term. The top-ranked documents should contain more query terms and assumed that indirectly related to a whole query. In this PRF-QE method, we select concept terms from terms within the PRF set based on the work [15] and finally select terms with the high specificity value based on collection and document frequency.

Given a set, $PRF = (x_1, x_2, ..., x_n)$ where x_i is a term in the PRF set. There are two conditions for finding concept terms; Equations 4 and Equation 5, respectively. The first condition measures how importance of term respect to the PRF set based on the distribution of term. The second condition measures the importance of term respect to entire collection.

$$p_{PRF}(x_i) = \frac{TF_{PRF}(x_i)}{PRFsize} > \theta_{c1} \tag{4}$$

and

$$\frac{p_{PRF}(x_i)}{p_{collection}(x_i)} > \theta_{c2} \tag{5}$$

where $TF_{PRF}(x)$ is a total frequency of term x in the PRF set and $PRFsize$ is a number of retrieved documents in PRF set. The $p_{collection}(x)$ is distribution of term x in the collection and measured by total collection frequency of term x respect to total documents in the collection.

To filter important concept, the entropy of a concept in the retrieved set is derived from the following equation.

$$G(x_i) = -p_{PRF}(x_i)\log(p_{PRF}(x_i)) \tag{6}$$

To select most related concepts, the conditional probability of them is used in the following conditions.

$$p_{PRF}(x_i|x_j) > \theta_{s_a} \quad \text{and} \quad p_{PRF}(x_j|x_i) > \theta_{s_b} \tag{7}$$

Now we have a number of concepts that dervied from Equation 4-7. Then finally select smaller number of concepts for expansion based on collection frequency and document frequency value as the following equation.

$$ConceptSpecificity(c) = \frac{log(CF(c))}{DF(c)} \tag{8}$$

where $CF(c)$ is the collection frequency of concept c and $DF(c)$ is document frequency that concept c appeared.

Cross Collection PRF-QE (CLEF-GPRF-QE). We assume that top-ranked documents that retrieved from external collection are also indirectly related with original query. Some queries are improved by this method as reported in work[8] and in our previous work [17]. In the current work we also use TREC Genomics 2004 collection as external source and then results from the initial retrieval in this collection are used to expand queries.

The method to select concepts is based on the PRF-QE method that described in the previous section. In addtion, threshold values of external collection are less than the CLEF collection, called target collection. Because the query is generated for the target collection therefore retreived size and related terms of the external are likely to be less than the target one.

Global Collection Query Expansion (Global-QE). We assign this method for the query with lowest $QPPpair$ value. This method is based on co-occurrence assumption that two terms frequently occur together in the same context, are likely to be related. The *pair* is two terms (t_i, t_j) within a small window length. This work fixes window length at 7 words for the co-occurrence condition. The set of pairs is $PairSet = (p_1, p_2, ..., p_k)$. Each pair p_i has the collection frequency c_i called *cotimes* and this value is used to select the most related of a query term. The cotimes set is $CoTime = (< p_1, c_1 >, < p_2, c_2 >, ..., < p_k, c_k >)$.

The steps of finding related terms from the co-occurrence terms are following.

1. Ascending sort the document frequency of query terms,
   $DFsorted = (q_1, q_2, ..., q_m)$.
2. Start from qi, lookup a corresponding pair(s) in the $PairSet$ then take them to the $CandidatePair$ set.
   – if q_i does not exits in the $PairSet$ set, increment i.
3. For each candidate pair, lookup the corresponding cotimes in the $CoTime$ set.
4. Descending sort candidate pairs according to the *cotimes* value,
   – if the pair p_i consists of one query term as (q_i, q_i), then cotimes c_i of the pair p_i is used to select other term(s),
   – otherwise, select top 5 terms from the sorted $CandidatePair$ set.

3.3 Readability based on Documents Complexity

To return most readability documents to user, we measure the complexity of document based on word length with the assumption that long word is the complex word. The document contains more complex words may be high complexity and uneasy to understand.

$$DocComplex(d) = \frac{TotalComplexWordInDocument(d)}{DocumentLength(d)} \tag{9}$$

After retreival we re-rank each 200 documents in the results set with the following equation.

$$ComplexVal(d) = 1 - 2.5 * DocComplex(d) \tag{10}$$

4 Experimental Design

We submit four runs where Run2 examines both query analysis method and query expansion method. Run3 and Run4 examine re-ranking method based on complexity of retrieved document.

Run1. The baseline retrieves with original title query using Lucene 4.6 [18] with StandardAnalyzer configuration and using Lucene's default VSM similarity.
Run2. The adaptive query expansion is using the $QPPpair$ to determine the expansion mechanism.
Run3. The re-ranking version of Run1 based on complexity of document.
Run4. The re-ranking version of Run2 based on complexity of document.

Determine Query Expansion Mechanisms. Queries are analyzed with the $QPPpair$ as described in Equation 1. Then each query is assigned with one of five group and detail as the shown in Table 1. We note that values of $QPPpair$ in each group are from human estimation.

Table 1. Query Expansion Mechanism based on QPPpair

Group	*QPPpairScore*	*QueryExpansionMechanism*
One	0.5 <= x	No expansion
Two	0.2 <= x < 0.5	PRF-QE with top 100 documents
Three	0.1 <= x < 0.2	CLEF-GPRF-QE with top 500 documents
Four	0 < x < 0.1	CLEF-GPRF-QE with top 1000 documents
Five	0 = x	Global-QE

Table 2. Parameter Values for Concept Finding in the Target Collection

Parameter	GroupTwo	GroupThree	GroupFour
TopK	100	500	1000
CollectionSize	914252	914252	914252
θ_{c1}	0.05	0.05	0.75
θ_{c2}	5.0	5.0	5.0
Entroy	0.05	0.05	0.00
θ_{sa}	0.08	0.08	0.08
θ_{sb}	0.03	0.03	0.03
MaxSelectTerm	7	9	9

Table 3. Parameter Values for Concept Finding in the External Collection

Parameter	GroupThree	GroupFour
TopK	500	1000
CollectionSize	3479789	3479789
θ_{c1}	0.025	0.025
θ_{c2}	2.5	2.5
Entroy	0.05	0.025
θ_{sa}	0.025	0.025
θ_{sb}	0.01	0.01
MaxSelectTerm	9	9
PartedSize	5	7

Adaptive Query Expansion. After determine group of query mechanism, we do expansion each group with the parameter settings as shown the Table 2 and Table 3. We also note that values of these parameters are from empirical trial.

5 Results

This task consists of two relevant judgments. The first one is traditional evaluation meaurement that shows retrieval performance such as P@10 and NDCG. The second measurement is readability-biased evaluation. The comparison of the P@10 with other teams shown in Fig1,2,3, and 4. The map, P@10,NDCG_cut_5 and NDCG_cut_10 of our four runs are shown in Table 4. The readability performance of all runs are shown in Table 5.

Table 4. The Retrieval Performance of Four Runs

Run	*map*	*p*@10	*ndcg_cut*_5	*ndcg_cut*_10
1	0.1090	0.2545	0.2354	0.2205
2	0.0930	0.2288	0.2047	0.1980
3	0.0219	0.0364	0.0248	0.0299
4	0.0180	0.0182	0.0169	0.0163

Table 5. The Readability-biasd Measure of All Runs

Run	$RBP(0.8)$	$uRBP(0.8)$	$uRPBgr(0.8)$
1	0.2785	0.2312	0.2251
2	0.2562	0.1818	0.1906
3	0.1679	0.1514	0.1425
4	0.0656	0.0600	0.0567

6 Discussion

Query Coherence Analysis. The query with more existing pairs in the corpus does not guarantee that the result set will be more relevant as seen that queries Q7,Q15,Q37, and Q51 have the high *QppPair* value but the P@10 value of these queries is not better than the median retrieval performance. Because Lucene's default similarity function computes weight of individual term not pair of terms.

Using the QPPpair to determine expansion mechanism is still better than expand all queries with the same mechanism. We believe that there are alternative ways to take advantage from these existing pairs such as reform the query to add more weight on the pair or search query with phrase option.

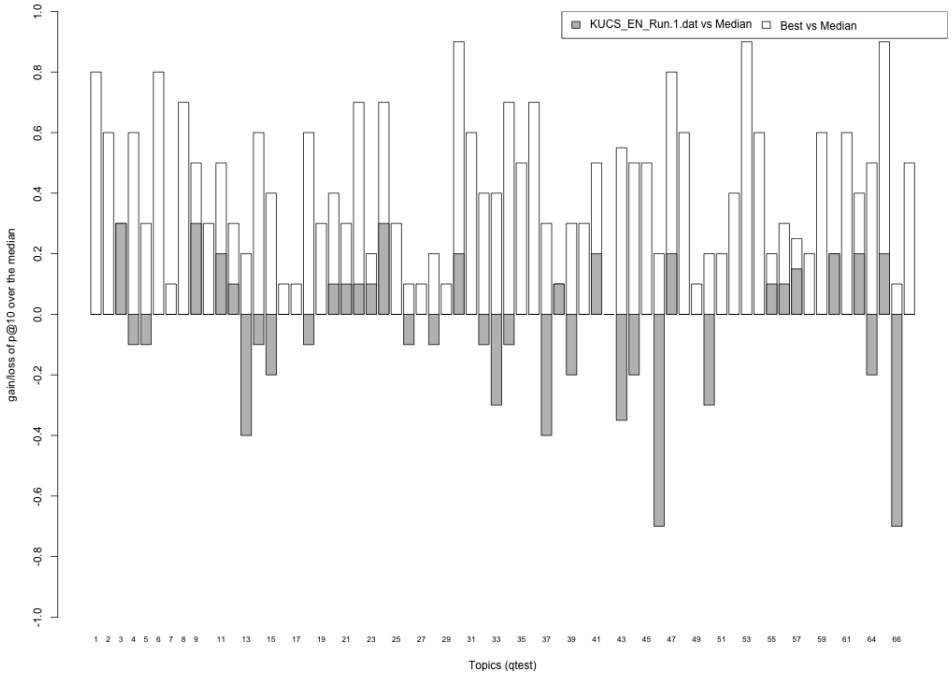

Fig. 1. Baseline Run1 compare with all teams

Query Expansion Mechanism. The work [15] have provided candidate terms for users to select in interactive manner but our work automatically selects terms from the same candidate set for expansion. Our expansion terms are mixing both useful and misleading terms because the theshold values as shown the Table 2 and Table 3 are setting with heuristic manner. By setting these parameters, we observe candidate terms derived from different threshold values of some queries. This is not a systematic manner for tuning parameter.

For expansion based on PRF set, the candidate terms sensitive to the tuning parameter. Even the queries are in the same group, the candidate terms for some query are likely to be related while others are drifting.

For global expansion method, the candidate terms are derived from the co-occurrence terms of the most specificity term in the query. This method is working for some query, for example, Q61. "fingernail bruises", the co-occurrence terms of "fingernail" are useful for expansion. On the other side, for example, Q59. "heavy and squeaky breath", the co-occurrence terms of the "squeaky" are not useful for expansion. This is the weakness of this method.

Document Complexity Method. The complexity document score is using only word length and then re-ranking retrieved documents. This method gets

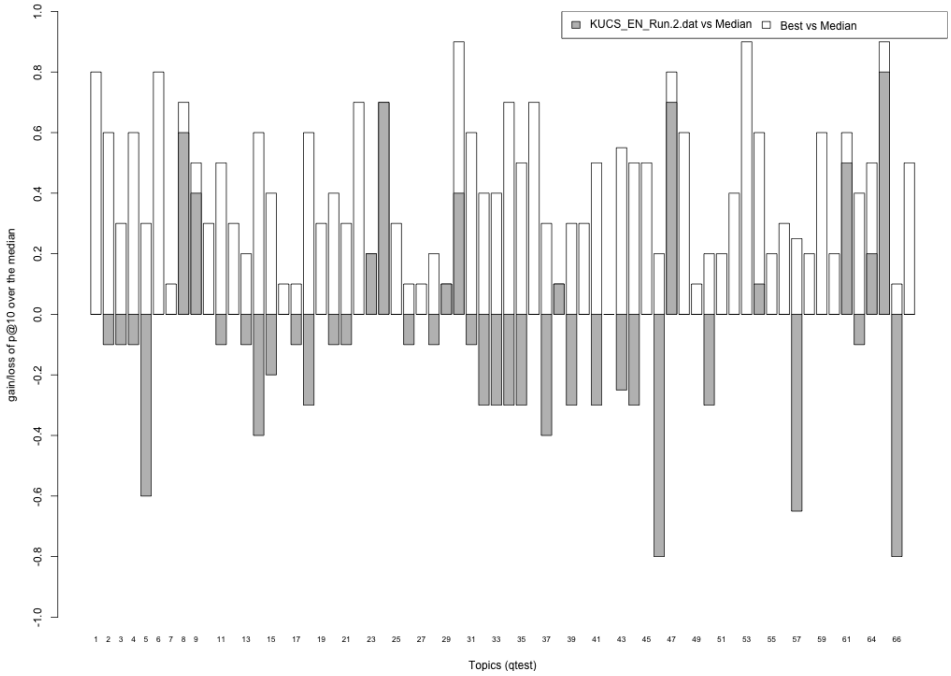

Fig. 2. Adaptive query expansion Run2 compare with all teams

the worse results. Therefore using more features of the document is necessary to estimate the complexity.

References

1. Lorraine Goeuriot, Liadh Kelly, Hanna Suominen, Leif Hanlen, Aurlie Nvol, Cyril Grouin, Joao Palotti, Guido Zuccon.: Overview of the CLEF eHealth Evaluation Lab 2015. CLEF 2015 - 6th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer, September (2015)
2. J. Palotti, G. Zuccon, L. Goeuriot, L. Kelly, A. Hanbury, G.J.F Jones, M. Lupu, and P. Pecina.: CLEF eHealth Evaluation Lab 2015, task 2: Retrieving Information about Medical Symptoms. In CLEF 2015 Online Working Notes. CEUR-WS (2015)
3. Task 2: User-Centred Health Information Retrieval, https://sites.google.com/site/clefehealth2015/task-2
4. Brigo, Francesco, et al.: Clearly written, easily comprehended? The readability of websites providing information on epilepsy. Epilepsy and Behavior 44, 35–39 (2015)
5. Vargas, Christina R., et al.: Readability of online patient resources for the operative treatment of breast cancer. Surgery 156(2), 311–318 (2014)
6. Sharma, Nikhil, Andreas Tridimas, and Paul R. Fitzsimmons.: A readability assessment of online stroke information. Journal of Stroke and Cerebrovascular Diseases 23(6), 1362–1367 (2014)

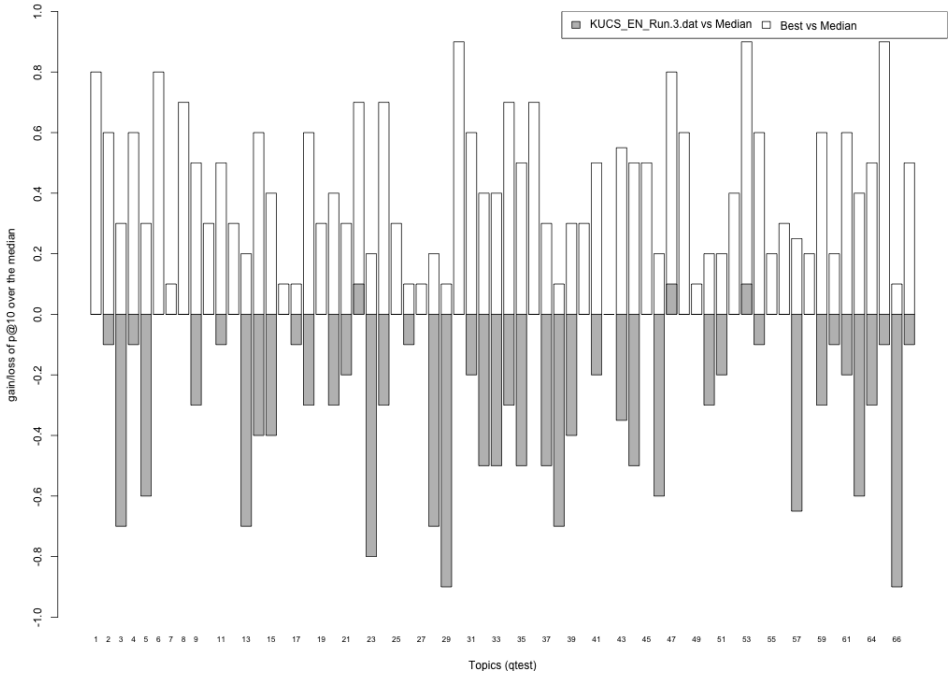

Fig. 3. Re-ranking of baseline Run3 compare with all teams

7. Cronen-Townsend, Steve, Yun Zhou, and W. Bruce Croft.: A framework for selective query expansion. In Proceedings of the thirteenth ACM international conference on Information and knowledge management, pp. 236–237, ACM (2004)
8. He, Ben, and Iadh Ounis.: Combining fields for query expansion and adaptive query expansion. Information processing and management 43(5) ,1294–1307 (2007)
9. Hauff, Claudia, Djoerd Hiemstra, and Franciska de Jong.: A survey of pre-retrieval query performance predictors. In Proceedings of the 17th ACM conference on Information and knowledge management, pp. 1419–1420, ACM (2008)
10. He, Ben, and Iadh Ounis.: Query performance prediction. Information Systems 31(7), 585–594 (2006)
11. Kumaran, Giridhar, and Vitor R. Carvalho.: Reducing long queries using query quality predictors. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 564–571, ACM (2009)
12. Bai, Jing, and Jian-Yun Nie.: Adapting information retrieval to query contexts. Information Processing and Management 44(6), 1901–1922 (2008)
13. Xu, Jinxi, and W. Bruce Croft.: Improving the effectiveness of information retrieval with local context analysis. ACM Transactions on Information Systems (TOIS) 18(1), 79–112 (2000)
14. Sanderson, Mark, and Bruce Croft.: Deriving concept hierarchies from text. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp 206–213, ACM (1999)

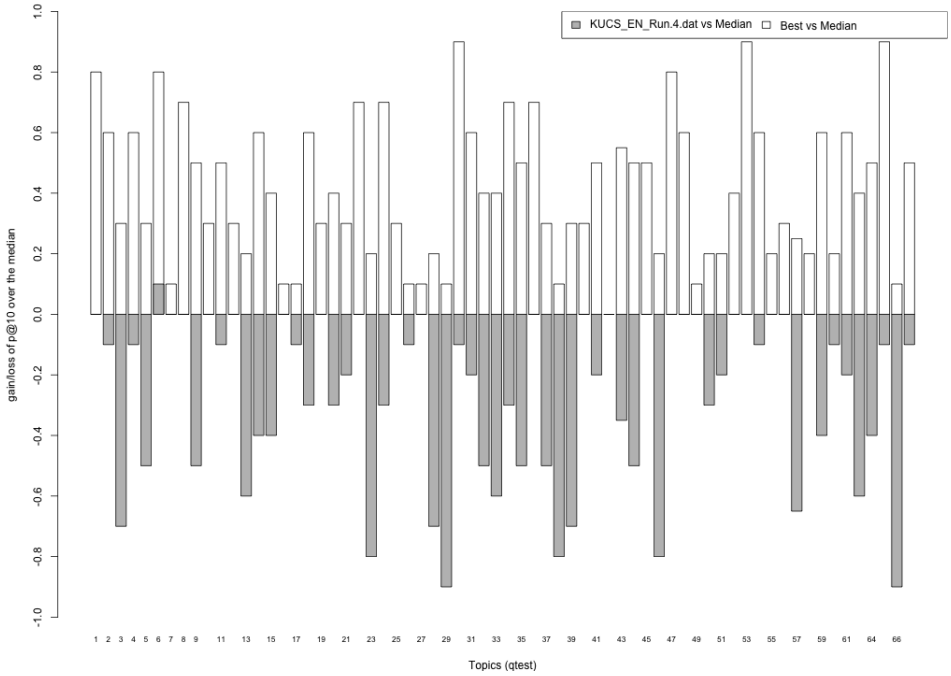

Fig. 4. Re-ranking of adaptive query expansion Run4 compare with all teams

15. Oh, J., Kim, T., Park, S., Yu, H., and Lee, Y. H.: Efficient semantic network construction with application to PubMed search. Knowledge-Based Systems. 39, 185–193 (2013)
16. Si, Luo, and Jamie Callan.: A statistical model for scientific readability. In Proceedings of the tenth international conference on Information and knowledge management, pp. 574–576, ACM (2001)
17. Thesprasith, Ornuma, and Chuleerat Jaruskulchai.: Csku gprf-qe for medical topic web retrieval. In Proceedings of the ShARe/CLEF eHealth Evaluation Lab (2014)
18. Apache Lucene, http://lucene.apache.org