Structural Query Expansion via motifs from Wikipedia

Joan Guisado-Gámez
DAMA-UPC
Universitat Politècnica de Catalunya
joan@ac.upc.edu

Arnau Prat-Pérez
DAMA-UPC
Universitat Politècnica de Catalunya
aprat@ac.upc.edu

Josep Lluís Larriba-Pey
DAMA-UPC
Universitat Politècnica de Catalunya
larri@ac.upc.edu

ABSTRACT

The search for relevant information can be very frustrating for users who, unintentionally, use inappropriate keywords to express their needs. Expansion techniques aim at transforming the users' queries by adding new terms, called expansion features, that better describe the real users' intent. We propose Structural Query Expansion (SQE), a method that relies on relevant structures found in knowledge bases (KBs) to extract the expansion features as opposed to the use of semantics. In the particular case of this paper, we use Wikipedia because it is probably the largest source of up-to-date information. SQE is capable of achieving more than 150% improvement over non-expanded queries and is able to identify the expansion features in less than 0.2 seconds in the worst-case scenario. SQE is designed as an orthogonal method that can be combined with other expansion techniques, such as pseudo-relevance feedback.

ACM Reference format:
Joan Guisado-Gámez, Arnau Prat-Pérez, and Josep Lluís Larriba-Pey. 2017. Structural Query Expansion via motifs from Wikipedia. In *Proceedings of ExploreDB'17, Chicago, IL, USA, May 14-19, 2017,* 6 pages.
DOI: http://dx.doi.org/10.1145/3077331.3077342

1 INTRODUCTION

Typically, users express their needs with queries consisting, usually, of a set of keywords. However, *vocabulary mismatch* between the keywords and the documents to be retrieved entails poor results that do not satisfy the user needs [12]. Poor results also arise from the *topic inexperience* of the users and, consequently, their tendency to use too general keywords [17].

Query expansion techniques aim at improving the results achieved by a user's query by means of introducing new expansion terms, called **expansion features**. Expansion features introduce new concepts that are semantically related with the concepts in the user's query and that allow overcoming part of the aforementioned problems. Different families of expansion techniques differ in the way they acquire the expansion features. One of such families consists in using knowledge bases (KBs). A KB consists of a set of linked entries, each of which describes a single concept, forming a graph, where the nodes represent the entries and the edges the relationships among them.

ExploreDB'17, Chicago, IL, USA

978-1-4503-4674-0/17/05…$15.00
DOI: http://dx.doi.org/10.1145/3077331.3077342

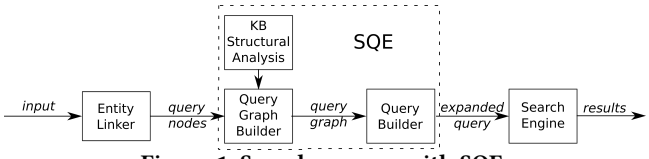


Figure 1: Search process with SQE.

We propose Structural Query Expansion (SQE), a new query expansion strategy that exploits KBs relying exclusively on the structural relationships of its entries. SQE is tightly linked to the KB that it exploits. Thus, as shown in Figure 1, it requires an offline analysis of its structural properties. These are used to define a set of motifs which allow building the **query graphs**. We define the query graph as the subgraph from the KB graph that contains the **query nodes**, which are the entries of the KB that represent the entities in the user's query, and the **expansion nodes**, out of which we extract the expansion features.

Although SQE can be used for any KB, in this paper, we use Wikipedia because it is the largest up-to-date source of information. However, its size represents a challenge because the search space in the Wikipedia graph is very large which creates problems related to finding efficiently meaningful relationships for the expansion process. The KB graph that we build from Wikipedia has two types of nodes that correspond to the Wikipedia's **articles** and **categories**. The edges of the graph are the Wikipedia hyperlinks, which connect articles and categories. The query nodes are always articles, since each Article represents a single topic. In this paper, we present a structural analysis of Wikipedia and we define a set of structural motifs that are capable of capturing reliable expansion features.

SQE obtains statistically significant improvements of more than 150% over the non-expanded queries by running in the order of a few tenths of a second at most. Additionally, as shown in Section 4, SQE is orthogonal to other expansion techniques. Particularly we show that combining SQE with pseudo-relevance feedback achieves up to 13.68% improvement in the quality of the results.

The contributions of this paper are summarized as follows:

- We propose SQE an expansion strategy that relies on KBs' structure. We implement it using Wikipedia.
- We propose SQE as orthogonal to existing strategies and it is executed in sub-second times.
- We test SQE with three different datasets and validate the results with statistical significance analysis.

2 STRUCTURAL QUERY EXPANSION

SQE consists of i) the structural analysis of the KB graph, ii) the query graph builder and iii) the query builder.

The structural analysis requires a ground truth that relates a set of queries with their optimal query graphs. The goal is to analyze

them to reveal their shared structural characteristics. This is an offline process that is done once.

The query graph builder materializes the revealed structural characteristics into a set of structural motifs in a way that, given a query node, we can infer its query graph. The goal is that the calculated query graphs have similar structural characteristics as those in the ground truth. In this paper, to validate our hypothesis, we have crafted the structural motifs empirically, to avoid introducing potential errors derived from an automatic algorithm.

SQE builds the expanded query with the user's query and the expansion features from the query nodes and the expansion nodes.

2.1 Wikipedia Structure Analysis

The analysis of the Wikipedia structure uses the optimal query graphs provided in a published ground truth [10] which relates each of the queries of the Image CLEF benchmark with its optimal query graph, i.e. the one made of the expansion nodes that allow achieving the best precision results. Although the ground truth is built using a particular query set, in Section 4 we show that the revealed characteristics are consistent for other query sets.

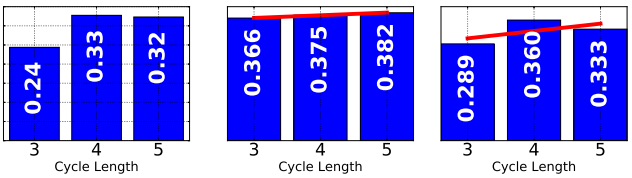
The analysis of the ground truth reveals that most of the expansion nodes of the optimal query graphs are connected to the query nodes by means of cycles of length 3, 4 and 5. Cycles are defined as a closed sequence of nodes, either articles or categories, with at least one edge among each pair of consecutive nodes. Actually, using the nodes of the cycles as source of expansion features allows achieving a precision of 0.833, 0.624, 0.588 and 0.547 for the top-1, top-5, top-10 and top-15 results respectively. This results are comparable to the best results achieved in the Image CLEF 2011 conference [15]. However, results in [15] were achieved combining textual and visual analysis techniques, using the three languages in which the metadata of the documents are written (while only English is used in this case), and exploiting feedback relevance techniques. The specific goal of the structural analysis for Wikipedia requires understanding which are the characteristics of these cycles in the optimal query graphs.

Figure 2a shows the average contribution of the cycles depending on their length. The contribution is a value between 0 and 1 that measures the part of precision that is obtained thanks to the cycles of a given length with respect to the one achieved by the whole query graph. We observe that the contribution of the cycles depending on their lengths is comparable to each other, although, according to the results, larger cycles seem to contribute more.

Regarding the observed proportion of articles and categories, Figure 2b shows the ratio of categories per cycle length. Approximately a third of the nodes are categories. Thus, categories play an important role maintaining the cycles within a single or very related domain of knowledge.

Finally, Figure 2c shows the average density of extra edges with respect to the length of the cycles. The density of extra edges is defined as the amount of edges minus the minimum required amount of edges to create a cycle divided by the maximum amount of possible edges of the cycle (two consecutive nodes can be connected by two edges). From Figures 2a and 2c, we can see a correlation between denser cycles and those that contribute more.

Summarizing the characteristics that let us differentiate good from bad cycles, we consent that:



(a) (b) (c)

Figure 2: Average (a) contribution (b) category ratio (c) density of extra edges; of cycles and their length.

- Cycles of length 3, 4 and 5 are to be trusted to reach articles that are strongly related with the query nodes.
- A third of the nodes of cycles have to be categories.
- The expansion features obtained through the articles of dense cycles are capable of leading to better results.

2.2 Query Graph Builder

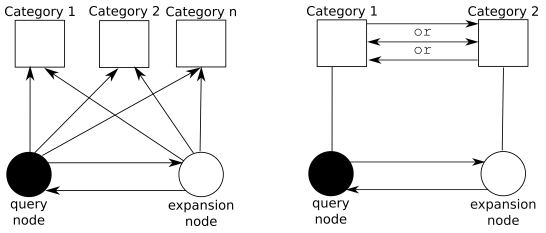
To calculate the query graphs in a way that they have the same structural characteristics as those in the ground truth, we propose the motifs depicted in Figures 3a and 3b, which are based on cycles of length 3 and 4 respectively. The motif depicted in Figure 3a is called, from now on, **triangular motif**, whereas the one depicted in Figure 3b is called **square motif**. In the figures, square nodes are categories, and round nodes are articles. Black round nodes are query nodes, while white ones are expansion nodes, which have been selected because they are part of a motif.

In the triangular motif, we force the query node to be doubly linked with the expansion node. That means that the query node actually links, in Wikipedia, to article (expansion node), and the article links, reciprocally, to the query node. Moreover, the article must belong to, at least, the same exact categories as the query node. In the square motif of Figure 3b, the query node and the new article must be also doubly linked. However, compared to the triangular motif, it is just required that at least one of the categories of the query node is inside one of the categories of the expansion node, or *vice versa*. Both patterns are chosen because these cycles fulfill the edge density and ratio of categories requirements. Note that we have avoided cycles of length 5 for performance reasons.

SQE consists in, given the query nodes as a starting point, identify all the nodes of the Wikipedia graph that are part of a motif and add them to the query graph. At the same time, while the motifs are being traversed, we build a set of pairs $< a, |m_a| >$, where a is an article that has appeared among the expansion nodes, and $|m_a|$ is the number of motifs in which it has appeared.

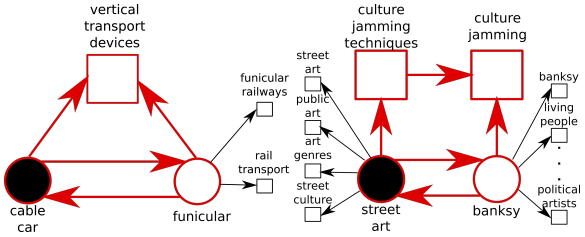
In Figure 4a we show an example of a triangular motif that for the Image CLEF query #93, "cable cars". Thanks to the motif, the article **funicular**, that is a similar transport system, becomes a part of the query graph. Similarly, in Figure 4b, for query #73, "graffiti street art on walls", the square motif introduces in its query graph the article **Banksy**, who is a graffiti artist.

2.2.1 Combining Query Graphs: SQE$_C$. We have designed a variation of the SQE which consists in combining the set of results instead of combining the set motifs. SQE$_C$ builds n different expanded queries, each using a particular motif configuration, each of which is used to retrieve the results. Finally, these sets are combined into a single one. Although in Section 4 we show in detail the proper configuration to maximize the performance overall analyzed tops,



(a) Triangular motif. (b) Square motif.

Figure 3: Expansion motifs.



(a) Triangular motif. (b) Square motif.

Figure 4: Expansion motifs in action for Image CLEF.

we anticipate some results: the triangular motif allows achieving better precision in small tops of results, up to five; while, the square motif allows achieving precision in large tops of results.

2.3 Query Builder

We first introduce the SQE retrieval model, which is based on a combination of the language modeling [13] and inference network [16]. The query likelihood model that we adopt is a factor between a multi-word query Q, and a document D represented as a bag of words as $P(Q|D) = \prod_{w_i \in Q} P(w_i|D)$. The feature function used to match words (document features), w to a document D is a Dirichlet smoothed probability: $P(w|D) = \frac{tf_{w,D}+\mu P(w|C)}{|D|+\mu}$, which generalizes to n-grams and unordered term proximity.

We build the expanded query as a three-part combination: i) the user's query, ii) the titles of the query nodes, and iii) the titles of the articles expansion nodes. Titles are taken as a n-gram of consecutive terms for phrase matching. In the expanded query, the expansion features are **weighted** proportionally to the number of motifs in which they have appeared, i.e. the title of a is weighted proportionally to $|m_a|$. Note that this means that we are also exploiting the structural properties to build the query.

3 EXPERIMENTAL SETUP

In this section we provide details to reproduce our experiments. Experiments described in this paper are implemented using Indri [14], an open source search engine.

The **entity linker** that we use is implemented using Dexter [3], which is an open source tool that recognizes entities in a given text and links them with Wikipedia articles, our query nodes. If Dexter is not able to find any matching entry, we preprocess the text using Alchemy [18], a tool that identifies entities but does not link them with Wikipedia. According to our experiments, the combination of both Dexter and Alchemy achieves more than 80% precision in identifying and linking the entities.

We have used **Image CLEF** to design SQE. The collection of results contains 237,434 images each of which have short descriptions as metadata, which we use as our target documents. Approximately, 60% of these descriptions contain texts in English. We have used **CHiC 2012 & CHiC 2013** to evaluate SQE. These datasets are based on cultural heritage retrieval. Both datasets shared the collection of results, which contains 1,107,176 short documents. Each of the three datasets provides a set of fifty queries (total 150) and their corresponding valid results.

We use the English Wikipedia dump of July 2nd, 2012 as our KB. It has 9,483,031 articles and 99,675,360 links among articles, 1,320,671 categories, 3,795,869 links among categories and 41,490,074 links among articles and categories.

To evaluate the results, we focus on the analysis of the system's precision for the default tops in TrecEval. To show the statistical significance with p<0,05, we have done the paired t-test.

	P@5	P@10	P @15	P@20	P@30	P@100	P@200	P@500	P@1000
QL_Q	0.136	0.130	0.121	0.112	0.089	0.035	0.018	0.007	0.003
QL_E	0.248	0.226	0.220	0.213	0.197	0.125	0.077	0.038	0.020
$QL_{Q\&E}$	0.244	0.220	0.213	0.210	0.195	0.127	0.081	0.040	0.021
SQE_T	0.456†	0.402†	0.384†	0.349†	0.282†	0.147†	0.0859†	0.040†	0.020†
$SQE_{T\&S}$	0.448†	0.414†	0.400†	0.379†	0.315†	0.171†	0.102†	0.048†	0.025†
SQE_S	0.444†	0.402†	0.387†	0.362†	0.301†	0.164†	0.104†	0.051†	0.027†
SQE^{UB}	0.578	0.519	0.494	0.485	0.382	0.188	0.117	0.054	0.028

Table 1: Comparison of the precision. † indicates statistically significant improvement.

4 EXPERIMENTS

We use the query likelihood (QL) model as state-of-the-art retrieval baseline and compare our technique (SQE) with the user's query (QL_Q), the query entities (QL_E) and the expansion features (QL_X).

4.1 SQE Configuration

First, we analyze SQE and compare the results with the ones achieved by the used ground truth. Then, we use this analysis to configure SQE_C, described in Section 2.2.1. For this purpose, we use ImageCLEF because we have its ground truth query graphs. For these experiments, we select manually the query entities to avoid any noise that could be introduced due to the errors of the entity linker. In more detail, we compare QL_Q, QL_E and $QL_{Q\&E}$ (which combines the user's query and the query entities) with SQE when only the triangular motif is used, SQE_T, when only the square motif is used, SQE_S and when the combination of both motifs is used to create query graphs, $SQE_{T\&S}$. Also, the ground truth is used to build and upper bound, SQE^{UB}.

In Table 1, we see that the SQE_T, $SQE_{T\&S}$ and SQE_S improve significantly the precision achieved by any of the baselines (QL_Q, QL_E and $QL_{Q\&E}$) for all tested levels of precision. This means that the achieved improvement is due to the introduction of the expansion features, and not only due to the query entities. Also, we see that the results achieved by SQE represent, in the worst-case scenario (SQE_S, P@20 - 0.362), the 71.41% of the upper bound results (SQE^{UB}, P@20 - 0.485). In average, this percentage is 85.86%, which means that the proposed query expansion strategy is close to the results achieved by the upper bound. Note that the results achieved by SQE^{UB} are due the use of ground truth query graphs. On the other hand, the results achieved by SQE traverse blindly the whole Wikipedia using the described motifs to create the query

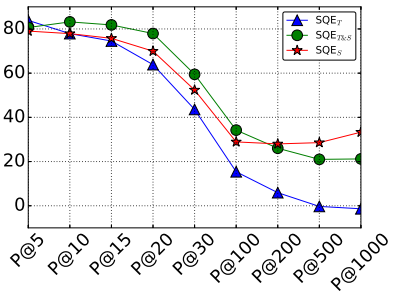


Figure 5: Percentage improvement over the maximum of QL_Q, QL_E and $QL_{Q\&E}$.

graphs. Thus, although these query graphs have the same structural properties as those in the ground truth, they are not equal.

In Figure 5 we show the percentage improvement of the three configurations of SQE with respect to the best result achieved by either QL_Q, QL_E and $QL_{Q\&E}$. We observe that the improvement diminishes as the size of the top increases. To understand this behavior, we need to look at that the average number of correct documents per query, which is 68.8. Hence, it is difficult to improve the precision when the amount of retrieved documents is much larger than the amount of actually valid documents. A deeper analysis of these configurations reveals three different ranges, depending on the query expansion configuration that achieves the best results. The first range includes up to P@5, the second range that from P@5 to P@100 and the third from P@100 to P@1000.

Range P@1-P@5: SQE_T achieves the best results, which introduces 0,76 expansion features in the query graph per user query. SQE_T achieves 83.87% improvement However, since there are just a few expansion features, the expanded query is not very different from the user's query, and the improvement decreases quickly.

Range P@5-P@100: The best results are achieved by $SQE_{T\&S}$. The improvement goes from 83.85% to 34.22%. This configuration introduces, in average, 20.96 expansion features per query. Combining expansion features very close to the original query, obtained via the triangular motif, with other no so close, vis square motifs, makes this configuration the best for this range in the middle.

Range P@100-P@1000: The configuration that allows achieving the best results is the SQE_S whose improvement ranges from 27.99% to 33.30%. It introduces, in average, 20.48 expansion features per query. The fact that these expansion features are not so tied to the original query issued by the user enables to retrieve documents that were not selected by the other configurations.

Hence, we have configured SQE_C combining the results achieved by the executions of SQE_T, $SQE_{T\&S}$ and SQE_S in a way that the first five results come from SQE_T, the next 195 results come from $SQE_{T\&S}$ and the rest of the results come from SQE_S.

4.2 SQE Evaluation

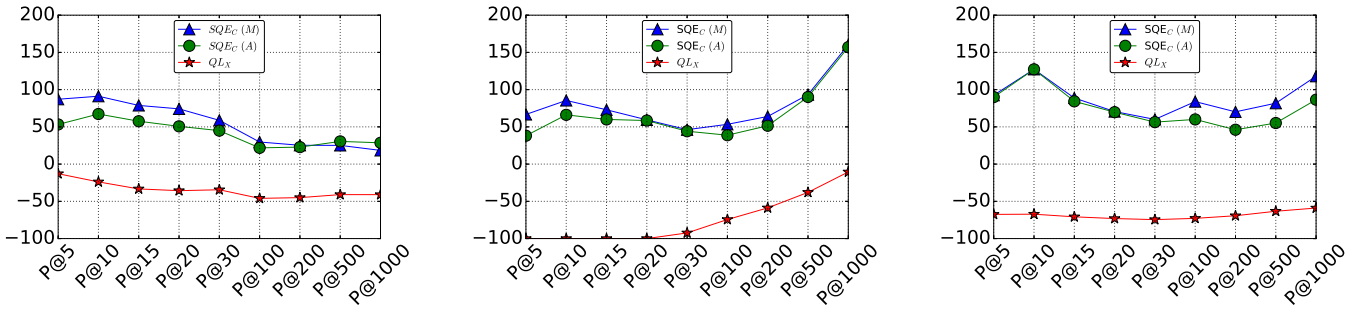
Now we evaluate the variation of SQE described in Section 2.2.1 as SQE_C and configured as described in the previous section. We use three datasets to test whether the results are consistent among them. In Figure 6 we show the percentage improvement achieved by SQE over the best execution for each top using QL_Q, QL_E and $QL_{Q\&E}$ configurations. Also, we use the percentage improvement of using only the expansion features (QL_X). Note that SQE_C (M) selects manually the query entities, whereas in SQE_C (A) are selected automatically by the entity linker described in Section 3. In Figure 6, we see that for the three datasets using the expansion features in an isolated way is not useful to improve the precision of the system, but diminishes the quality of the results. That supports the idea of assembling the expanded query as described in subsection 2.3. The user's query, even not being the best way to express the real intention of the user, due to his/her lack of knowledge and the vocabulary mismatch, is the only query form in which we are sure that the system has not introduced any error and hence, it helps to diminish errors that could be introduced later in the process. The query entities reinforce the user's query removing all the signs of ambiguity from the user's intent. Finally, the expansion features are helpful to overcome the classical problems of information retrieval.

We observe that SQE, which is depicted as SQE_C (M) and SQE_C (A), improves the results significantly for all datasets. We also observe that there are differences between selecting the entities manually or automatically. The manual entity selection is almost an upper bound of SQE because it isolates the creation of the query graphs from errors that could be introduced due to the entity linking module. Nonetheless, we observe that in the worst-case scenario (Image CLEF, P@5), the improvement achieved by SQE_C (A) represents 81.89% of the result achieved by SQE_C (M). As shown in Figure 6c, there is also a difference between the results achieved by SQE_C (M) and SQE_C (A) for the larger tops, while in small tops is imperceptible. It is difficult to explain why in Entity linking is not the focus of this paper, however, improving the techniques used in our system would improve the results, making it possible to achieve the results of selecting manually the entities and the query nodes.

In Tables 2a, 2b and 2c, we show the precision achieved for the three datasets. In particular, we show the results achieved by our baselines, the expansion features and SQE. The results show that both SQE_C (M) and SQE_C (A) present statistically significant improvements with respect to the baselines for the three datasets.

Note that combining query graphs allows, in Table 2a, achieves better results than each of the configurations independently in their best range, in Table 1 This supports our strategy of combining the results obtained by different expanded queries to improve the quality of the results independently the amount of them.

Focusing only in SQE_C (A) we also observe differences among the results. A superficial analysis could induce us to think that it performs better for Image CLEF because the precision achieved with this dataset goes from 0.380 (P@5) to 0.029 (P@1000), while for CHiC 2012 and 2013 it goes from 0.232 to 0.013 and from 0.304 to 0.017 respectively. We could also think this could be due to an overfitting of SQE for Image CLEF, since it is the training dataset. However, there are objective facts that explain this behavior. First, the document collection of Image CLEF consists of 237,434 documents, while the document collection of the CHiC datasets has 1,107,176. This makes Image CLEF an easier dataset. Moreover, Image CLEF has an average of 68.8 correct results per query, while CHiC 2012 and CHiC 2013 have 31.32 and 50.6 respectively. In addition, all the queries in Image CLEF have at least 1 correct result, while in CHiC 2012 there are 14 queries (out of 50) that do not have any correct results and in CHiC 2013 there is 1 query without any correct result. Note that the larger the number of valid results



(a) Image CLEF (b) CHiC 2012 (c) CHiC 2013

Figure 6: Percentage improvement of the query expansions selecting the entities manually SQE_C (M) and automatically SQE_C (A) and also of the expansion features isolatedly.

	P@5	P@10	P @15	P@20	P_@30	P@100	P@200	P@500	P@1000
QL_Q	0.136	0.130	0.121	0.112	0.089	0.035	0.018	0.007	0.003
QL_E (M)	0.248	0.226	0.220	0.213	0.197	0.125	0.077	0.038	0.020
QL_E (A)	0.156	0.134	0.145	0.147	0.137	0.107	0.069	0.035	0.022
$QL_{Q\&E}$ (M)	0.244	0.220	0.213	0.210	0.195	0.127	0.081	0.040	0.021
$QL_{Q\&E}$ (A)	0.148	0.124	0.133	0.138	0.133	0.107	0.069	0.035	0.022
Q_X	0.216	0.172	0.147	0.137	0.129	0.069	0.045	0.023	0.013
SQE$_C$ (M)	0.464†	0.432†	0.393†	0.371†	0.313†	0.165†	0.102†	0.050†	0.027†
SQE$_C$ (A)	0.380†	0.378†	0.347†	0.321†	0.286†	0.155†	0.100†	0.052†	0.029†

(a) Image CLEF results.

	P@5	P@10	P @15	P@20	P_@30	P@100	P@200	P@500	P@1000
QL_Q	0.148	0.100	0.084	0.077	0.074	0.034	0.018	0.007	0.004
QL_E (M)	0.156	0.118	0.108	0.101	0.093	0.042	0.023	0.010	0.005
QL_E (A)	0.100	0.072	0.067	0.061	0.053	0.021	0.011	0.007	0.004
$QL_{Q\&E}$ (M)	0.168	0.124	0.113	0.106	0.097	0.044	0.023	0.010	0.005
$QL_{Q\&E}$ (A)	0.116	0.086	0.076	0.068	0.057	0.022	0.012	0.007	0.004
Q_X	0.000	0.000	0.000	0.000	0.007	0.011	0.010	0.006	0.005
SQE$_C$ (M)	0.280†	0.230†	0.196†	0.169†	0.141†	0.067†	0.038†	0.020†	0.013†
SQE$_C$ (A)	0.232†	0.206†	0.181†	0.168†	0.139†	0.061†	0.035†	0.019†	0.013†

(b) CHiC 2012 results.

	P@5	P@10	P @15	P@20	P_@30	P@100	P@200	P@500	P@1000
QL_Q	0.160	0.110	0.101	0.092	0.084	0.045	0.028	0.011	0.006
QL_E (M)	0.132	0.110	0.119	0.115	0.104	0.054	0.035	0.016	0.009
QL_E (A)	0.104	0.078	0.076	0.065	0.058	0.034	0.026	0.015	0.008
$QL_{Q\&E}$ (M)	0.132	0.110	0.119	0.119	0.110	0.056	0.036	0.017	0.009
$QL_{Q\&E}$ (A)	0.104	0.082	0.080	0.071	0.062	0.035	0.026	0.015	0.008
Q_X	0.052	0.036	0.035	0.032	0.028	0.015	0.011	0.006	0.004
SQE$_C$ (M)	0.308†	0.250†	0.224†	0.203†	0.176†	0.103†	0.062†	0.030	0.020†
SQE$_C$ (A)	0.304†	0.250†	0.219†	0.202†	0.172†	0.090†	0.053†	0.026†	0.017†

(c) CHiC 2013 results.

Table 2: Comparison of the precision. † indicates statistically significant improvement.

for the queries, the easier it is to resolve them. Hence, the highest precision is achieved for the Image CLEF collection, then comes CHiC 2013 and finally, CHiC 2012. Moreover, we observe that the percentage improvement, shown in Figure 6, for the three datasets is equivalent, and even better for the CHiC 2013 dataset.

4.3 Pseudo-Relevance Feedback comparison

Now we compare SQE with pseudo-relevance feedback (PRF), a state-of-the-art expansion model which extract the expansion features from the top documents retrieved by the query. We use PRF as an adaptation of Lavrenko's relevance model [8]. In this model, the original query Q is used to retrieve a ranked list of documents D ordered by $P(Q|D)$ and sort their concepts by $P(w|Q)$ to keep top n concepts, which are the expansion features. Then it combines the original query with the expansion features. The relevance model, $P(w|Q)$, is computed as: $P(w|Q) = \frac{\sum_D (P(w|D)P(Q|D)P(D)}{P(Q)}$.

	P@5	%G	P@10	%G	P@15	%G	P@20	%G	P@30	%G
PRF_Q	0.000	-100	0.000	-100	0.000	-100	0.001	-99.11	0.000	-99.22
PRF_E	0.004	-97.44	0.004	-91.01	0.004	-97.25	0.003	-97.96	0.002	-98.54
$PRF_{Q\&E}$	0.004	-97.30	0.002	-98.39	0.003	-97.98	0.003	-97.83	0.003	-97.96
SQE_C/ PRF	0.432	+13.68	0.370	-2.12	0.348	+0.39	0.323	+0.62	0.289	+1.15

(a) Image CLEF results.

	P@5	%G	P@10	%G	P@15	%G	P@20	%G	P@30	%G
PRF_Q	0.000	-100	0.002	-98.00	0.001	-98.45	0.001	-98.70	0.000	-99.22
PRF_E	0.000	-100	0.008	-88.89	0.004	-92.05	0.005	-91.80	0.005	-98.54
$PRF_{Q\&E}$	0.000	-100	0.004	-95.35	0.003	-96.45	0.002	-97.06	0.001	-97.96
SQE_C/ PRF	0.244	+5.17	0.218	+5.83	0.193	+6.60	0.173	+2.98	0.145	+3.85

(b) CHiC 2012 results.

	P@5	%G	P@10	%G	P@15	%G	P@20	%G	P@30	%G
PRF_Q	0.000	-100	0.004	-96.36	0.004	-96.05	0.003	-96.74	0.003	-96.07
PRF_E	0.000	-100	0.008	-89.74	0.007	-91.18	0.008	-87.69	0.007	-88.45
$PRF_{Q\&E}$	0.004	-96.15	0.006	-92.68	0.005	-93.38	0.006	-91.55	0.005	-91.45
SQE_C/ PRF	0.288	-5.26	0.264	+5.60	0.237	+8.52	0.220	+8.91	0.193	+12.39

(c) CHiC 2013 results.

Table 3: Precision achieved using PRF. "%G" stands for percentage gain with respect to precision in Table 2.

For the three datasets, in Table 3 we show the results achieved using PRF with the user's query (PRF_Q), with the query entities (PRF_E) and both ($PRF_{Q\&E}$). We also show the percentage gain with respect to the QL_Q, QL_E and $QL_{Q\&E}$ in Table 2. These results show that PRF is not particular useful to improve the precision in any of the analyzed tops. On the contrary, PRF seems to worsen the results. Although PRF has proven to be a useful expansion model allowing to achieve relevant improvements for many queries, it does not allow identifying good expansion features for the tested datasets. Also, we show, for the three datasets, the results achieved by combining SQE with PRF. In this case SQE is used to generate a query, then this query is used by PRF as previously described to reformulate the query and retrieve the documents. We observe that in most of the analyzed tops the combination of PRF with SQE allows improving the results shown previously in Table 2.

Note that although PRF techniques do not improve the results of **non**-expanded queries of our datasets, it uses and benefits from the SQE expansion. Actually, SQE is designed to be orthogonal to many other techniques some of them reviewed in Section 5.

4.4 SQE Performance

We used an Intel Xeon CPU E5-2609 with 128GB of RAM. We have not used any technique, such as indexing or exploiting parallelism,

	Image CLEF	CHiC 2012	CHiC 2013
SQE$_T$	47.06	74.09	51.80
SQE$_{T\&S}$	94.12	178.20	119.84
SQE$_S$	52.42	105.94	69.18
Total Time	1373.38	8907.76	5361.34

Table 4: Execution times in milliseconds.

to speed up the process. In Table 4 we show the time spent generating the query graphs by means of using the described motifs: the triangular, the square and the combination of both. Note that the total query expansion time is negligible compared with the whole process. In the worst-case scenario, which is the Image CLEF dataset, the time spent building the three query graphs represents 14% of the total, while in the two other datasets this only represents 4%. Also, the maximum amount of time spent building the query graphs, 358.23 for the CHiC 2012 collection, is not burden for real-time systems. This time would probably be easily reduced by parallelizing the expansion process.

5 RELATED WORK

Wikipedia has become a frequently used source of expansion features Egozi et al. [5] use it to rewrite the queries by using PRF. This technique depends on the quality of the pseudo-relevance feedback expansion, which is very poor in our query set unless we previously expand the queries. In [2], the expansion features are extracted out of the most important terms of the Wikipedia articles and are calculated with classical TF-IDF. Wikipedia is also used to derive search support tools. For instance, in [11], Wikipedia is used to build a map of concepts, then, users' queries are mapped onto those concepts, which make it easier for the search engine to resolve them. In [6], the authors show that anchor texts are similar to real queries regarding to term distribution and length, therefore they can be a source of expansion features. This approach is interesting for us because we could use the anchor texts of the expansion nodes as another source of expansion features. Also, in [7], they build a virtual query log, that can be used to reformulate the queries. However, to the best of our knowledge, few techniques explore the structure of the KB, and those that do it, limit its use to the link level, ignoring more complex, and according to our results, relevant structures. For example, in [1], the authors propose a query expansion method for blog recommendation. Their method is based on the analysis of links. The anchor text of the most important twenty links is used to expand the query which results in a significant improvement in terms of precision. Such an approach could be used in our work to rate the importance of the links, and then, include the strength of connections in the motifs. Also, in [4], the authors contribute to the field with a relevant work consisting in exploiting the entities as source of expansion features. The authors propose a framework which builds a model in which the entities of documents and queries are central for the retrieval process. This approach is interesting, and we would like to explore it further. In contrast to previous works, in this paper we do not focus on analyzing the content of the KB, or ranking the links among articles using other techniques. In [9], the authors presented a proof of concept in which they used the network structure of a KB. Although they achieved good results, they borrow a metric for community detection in social networks and hence, they do not exploit the particular KBs structures.

6 CONCLUSIONS AND FUTURE WORK

We have proposed SQE as a query expansion technique that relies on the underlying network structures of KBs. SQE is a three-steps process. First, analyzes the KB structure to reveal relevant characteristics. Second, it materializes these characteristics into a set of motifs, which are used to relate the user's queries with a set of semantically connected entries from the KB with no need of semantic analysis. Third, it builds the expanded query with the original query and the expansion features extracted from the relevant entries.

From the analysis of Wikipedia, we have defined 2 different types of motifs: the triangular motif and the square motif. To evaluate SQE we have used three different datasets, Image CLEF, CHiC 2012 and CHiC 2013. The results achieved by SQE are consistent for the three datasets, thus that SQE is not overfitted for a particular one. From the results, we see that the triangular motif is useful to improve the results of small tops up to 83.87%, a combination of the triangular and the square motifs improve the result in between small and large tops up to 33.30%, while using the square motif exclusively improves the results of large tops up to 83.85%. Also, we have presented a way of combining several query graphs to improve the most no matter the top to be optimized.

We have succeeded in identifying the proper motifs for Wikipedia, however there are many KBs and probably each has its own relevant structures. We need to expand our understanding of KBs, and study what other motifs may be relevant for other KBs besides Wikipedia. For that purpose, we are already working on a learning algorithm that is capable of identifying such motifs automatically.

ACKNOWLEDGMENTS

DAMA-UPC thanks the Ministry of Economy, Industry and Competitiveness of Spain and Generalitat de Catalunya, for grant numbers TIN2013-47008-R and SGR-1187 respectively, also the EU H2020 for funding the Uniserver project (ICT-04-2015-688540). Also, thanks to Oracle Labs for the support to our research on graph technologies.

REFERENCES

[1] J. Arguello, J. Elsas, J. Callan, and J. Carbonell. 2008. Document Representation and Query Expansion Models for Blog Recommendation. In *ICWSM*.
[2] M. Boer, K. Schutte, and K. Wessel. 2015. Knowledge based query expansion in complex multimedia event detection. *MTA* (2015), 1–19.
[3] D. Ceccarelli, C. Lucchese, S. Orlando, R. Perego, and S. Trani. 2013. Dexter: an open source framework for entity linking. In *ESAIR*. 17–20.
[4] J. Dalton, L. Dietz, and J. Allan. 2014. Entity query feature expansion using knowledge base links. In *SIGIR*. 365–374.
[5] O. Egozi, S. Markovitch, and E. Gabrilovich. 2011. Concept-Based Information Retrieval Using Explicit Semantic Analysis. *TOIS* 29, 2 (2011), 8.
[6] N. Eiron et al. 2003. Analysis of anchor text for web search. In *SIGIR*. 459–460.
[7] Van Dang et. al. 2010. Query reformulation using anchor text. In *WSDM*. 41–50.
[8] V. Lavrenko et al. 2001. Relevance-Based Language Models. In *SIGIR*. 120–127.
[9] J. Guisado-Gámez, D. Dominguez-Sal, and J. Larriba-Pey. 2014. Massive Query Expansion by Exploiting Graph Knowledge Bases for Image Retrieval. In *ICMR*.
[10] J. Guisado-Gámez and A. Prat-Pérez. 2015. Understanding Graph Structure of Wikipedia for Query Expansion. In *GRADES*. 6:1–6:6.
[11] J. Hu, G. Wang, F. Lochovsky, J. Sun, and Z. Chen. 2009. Understanding user's query intent with wikipedia. In *WWW*. 471–480.
[12] D. Metzler, S.T. Dumais, and C. Meek. 2007. Similarity Measures for Short Segments of Text. In *ECIR*. 16–27.
[13] J. Ponte and W. Croft. 1998. A Language Modeling Approach to Information Retrieval. In *SIGIR*. 275–281.
[14] T. Strohman, D. Metzler, H. Turtle, and W.B. Croft. 2005. Indri: A language model-based search engine for complex queries. In *ICOIA*, Vol. 2. 2–6.
[15] T. Tsikrika, A. Popescu, and J. Kludas. 2011. Overview of the Wikipedia Image Retrieval Task at ImageCLEF. In *CLEF*.
[16] H. Turtle and W. Croft. 1991. Evaluation of an Inference Network-Based Retrieval Model. *TIS* 9, 3 (1991), 187–222.
[17] X. Wang, A. McCallum, and X. Wei. 2007. Topical N-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval. In *ICDM*. 697–702.
[18] D. Wolfram, A. Spink, B. Jansen, and T. Saracevic. 2001. Vox populi: The public searching of the web. *JASIST* (2001), 1073–1074.