Novel Query Suggestions

Initial Work Report

Ilona Nawrot*
Normandy University
GREYC, HULTECH
Campus Côte de Nacre
F-14032 Caen, France
ilona.nawrot@unicaen.fr

Oskar Gross
University of Helsinki
Department of Computer
Science and HIIT
Gustaf Hällströmin katu 2b
FI-00014 Helsinki, Finland
ogross@cs.helsinki.fi

Antoine Doucet
Université de La Rochelle
L3i Laboratory
Avenue Michel Crépeau
17042 La Rochelle, France
antoine.doucet@univ-lr.fr

Hannu Toivonen
University of Helsinki
Department of Computer
Science and HIIT
Gustaf Hällströmin katu 2b
FI-00014 Helsinki, Finland
hannu.toivonen@cs.helsinki.fi

ABSTRACT

Query auto-completion (QAC) is one of the most recognizable and widely used services of modern search engines. Its goal is to assist a user in the process of query formulation.

Current QAC systems are mainly reactive. They respond to the present request using past knowledge. Specifically, they mostly rely on query logs analysis [11, 10, 12] or corpus terms co-occurrences [8] and rank suggestions according to their similarity with the partial user query, their past popularity, or their temporal dynamics features (e.g. trends, bursts, seasonality in query popularity) [9]. Consequently, a suggestion to be recommended by the QAC system must be preceded with a substantial users' interest and *ipso facto* must be an old information. However, a growing amount of people turns to search engines to find novel information, that is emergent or recently created (not redundant) one. Conventional QAC systems are thus unable to fulfill the increasingly real-time needs of the users.

In this work-in-progress report, we introduce a new approach to QAC — the system filtering out potentially novel information and proactively delivering it to the users. It aims at providing the users with some novel insight. Thus, it caters for their open-ended or persistent and increasingly real-time information needs. The preliminary method proposed in this paper to evaluate this approach forms time specific suggestions based on a comparison of two corpora constantly being updated with new data from chosen sources. An unsupervised and language-independent algorithm relying on relative novelty of terms co-occurrences is used to generate suggestions. The initial experimental results demonstrate the effectiveness of the approach in recommending queries leading to novel information. Therefore, they prove that such a system can enhance the exploratory power of a search engine and support the proactive information search.

Categories and Subject Descriptors

H.5.2 [**Information retrieval**]: Retrieval models and ranking—*Novelty in information retrieval*; H.5.2 [**Information retrieval**]: Information retrieval query processing—*Query suggestion*

General Terms

Algorithms, Experimentation, Measurement

Keywords

Novel query suggestions, novel associations graph, query auto-completion, novelty, terms co-occurrences

1. INTRODUCTION

The notion of novelty is used in two different ways within the information retrieval field. It can either refer to newness, originality or to diversity (covering multiple different meanings or usage variants) [3]. In this paper the novelty or novel information is defined as the information possessing the quality of being objectively new or emergent, it means non-existent before. It must also possess an original element where its originality stems from having recently come into existence and being of a kind not seen before, as opposed to just representing less common or rare interpretations or

*Ilona Nawrot is also associated with Poznań University of Economics, al. Niepodległości 10, 61-875 Poznań, Poland.

Web-KR'14, November 3, 2014, Shanghai, China.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-1606-4/14/11 ...$15.00.
http://dx.doi.org/10.1145/2663792.2663799 .

meanings[1]. For instance, if a new hurricane named *Sally* was about to affect the eastern coast of the United States the term *Sally* would be considered a novel query suggestion for the query term *hurricane* as opposed to e.g. *Sandy* which was one of the largest Atlantic hurricanes on record.

Novel information usually presents a potential threat or opportunity and loses its value as time between its publication and acquisition passes. Thus, efficient filtering it out from the multitude of irrelevant or relevant but redundant information has always been of interest, not only in business or political circles but also in everyday life of ordinary people. Nowadays, due to constantly changing world, accelerating pace of life and ever growing amount of new information it gains in importance and in some domains becomes critical. Simultaneously, the amount of information exceeding human perception capabilities force people to rely on machines and specialized software, particularly on search engines, to satisfy their information needs.

Some initial work on incorporation of real-time relevant information from social network sites or news into search results has been recently proposed [1, 2]. However, to the best of our knowledge there was no previous research considering the introduction of novelty into search results and particularly into query suggestions. In this paper we define a new approach for query auto-completion (QAC) – proactive, novel query recommendations that directly addresses this challenge. Specifically, we propose a graph based algorithm that builds upon the statistical analysis of terms co-occurrences within sentences and their relative novelty. Then, we leverage the knowledge originating from the graph structure and novelty scores to rank potential query terminations and deliver those to users in the form of query suggestion. The initial experiments provide promising results showing the effectiveness of our method in forming query suggestions that can lead to novel information.

The rest of the paper is organized as follows. First, we present proactive, novel query recommendations method in Section 2. Then, in Section 3, we detail the results of our initial experiments. Finally, in Section 4 we conclude the paper with short discussion and future work directions.

2. METHOD

The proactive, novel query recommendations method is comprised of two phases:

1. Novel information detection.
2. Selection of candidate terms for query termination.

The first phase consists in finding novel terms associations in the corpus of newly acquired documents what is modeled using the two corpora comparison paradigm. The second one subsists in usage of those novel terms associations in order to generate query suggestions. The phases are described in the following subsections.

2.1 Novelty Detection

The identification of novel information requires a point of reference, it is a norm against which we could compare – an approximation of established knowledge. Thus we model this task as two corpora comparison paradigm in which a relatively small corpus of newly acquired documents (*sample corpus*) is compared with larger corpus representing previous knowledge (*normative corpus*)[2].

In the search engines case, the normative corpus can be usually simulated using data obtained from their indices, whereas sample corpus can be composed based on data from newly crawled documents or a part of those. For instance, only news, blogs or authoritative Web pages documents can be chosen to constitute a sample corpus. Specific implementation details will however depend on the application. In special case the normative corpus can be initially empty denoting no memory or no previous knowledge. Nevertheless, in general the bigger the normative corpus the better the performance (if the sample size increases then the power of any statistical test applied on it should also increase as the standard deviation should decrease).

Furthermore, the vocabulary and syntactic rules are relatively stable. Moreover, even though theoretically each information can be conveyed in an infinite number of ways, in each language there exist certain standards specifying what is correct and sounds well from the linguistic point of view. Taking both into consideration, it can be assumed that there is a limited number of preferred ways an information can be conveyed in short time intervals. As so, the vocabulary and terms associations (modeled using normative corpus) are relatively stable in short time intervals. Thus any new term, new term association or different frequency of usage of the given terms association in the sample corpus in comparison to the normative corpus can indicate novel information.

Consequently, given the two corpora novel information identification problem is then reduced to the problem of detecting outstanding terms associations present in the sample corpus in comparison to the normative one. This is accomplished by using the odds ratio (OR) — measure of association which additionally estimates the relative strength of such terms associations allowing their ranking according to novelty [6, 7].

Formally, let $n(t_i, t_j)$ denote the number of sentences in the sample corpus containing both terms t_i and t_j, whereas $m(t_i, t_j)$ its equivalent number in the normative corpus. Further, let N and M be the total number of sentences in the sample and normative corpus respectively. The relationship between the corpus type (sample or normative) and the incidence of terms pair (t_i, t_j) is presented in Table 1.

Table 1: Contingency table for the odds ratio test

	Sample corpus	Normative corpus
Sentences with target terms pair	$n(t_i, t_j)$	$m(t_i, t_j)$
Other sentences	$N - n(t_i, t_j)$	$M - m(t_i, t_j)$
Total	N	M

The odds of the given terms pair (t_i, t_j) occurring in the sample corpus is then defined as:

$$odds_S = \frac{n(t_i, t_j)}{N - n(t_i, t_j)}. \tag{1}$$

[1] Thus, novel information should not be identified with long-tail information as this still can be redundant and does not always possess the quality of being fresh or recent.

[2] Similarly, most people when asked to evaluate the novelty of a given information firstly analyze the given information and then compare it to their acquired knowledge (or to the representation of knowledge provided by the experimenter when asked to evaluate it in a hypothetical scenario).

Analogically, the odds of the terms pair (t_i, t_j) occurring in the normative corpus is defined as:

$$odds_N = \frac{m(t_i, t_j)}{M - m(t_i, t_j)}. \tag{2}$$

Finally, the odds ratio (OR) is given by the formula:

$$\begin{aligned} OR = \frac{odds_S}{odds_N} = \frac{\frac{n(t_i,t_j)}{N-n(t_i,t_j)}}{\frac{m(t_i,t_j)}{M-m(t_i,t_j)}} = \\ = \frac{n(t_i, t_j)\,(M - m(t_i, t_j))}{m(t_i, t_j)\,(N - n(t_i, t_j))}. \end{aligned} \tag{3}$$

If $OR = 1$ then the likelihood of using a given terms association in the sample corpus is equal to that of the normative one. When $OR > 1$ it means that a given terms association is more likely to occur in the sample corpus that in the normative one which can be a sign of novelty. Novel terms associations are then the ones satisfying the condition:

$$OR = \frac{odds_S}{odds_N} = \frac{n(t_i, t_j)\,(M - m(t_i, t_j))}{m(t_i, t_j)\,(N - n(t_i, t_j))} > 1. \tag{4}$$

2.2 Generation of Query Terminations

To leverage the knowledge originating from the structure of novel terms associations and their novelty scores the graph-based algorithm is proposed to select query terminations. Specifically, the query suggestion terms are chosen based on the *novel associations graph* (G_{NA}). This structure is obtained from novel terms co-occurrences identified in the previous (novelty detection) phase. It is defined as a pair (G_{NA}, w) where $G_{NA} = (V, E)$ is a graph such that V denotes the set of all novel terms, $E \subseteq V \times V$ their associations and $w : E \to (1; \infty)$ is a weight function assigning each edge its novelty score (OR value; see Equations (1–4)). Given the first query term and the novel associations graph G_{NA} the query terminations are then generated in two steps: preparation and selection.

Preparation.

The novel information detection phase does not account for the senses [13] in which the terms co-occur. Hence, higher-level neighbors of the given query term in the novel associations graph G_{NA} might not be semantically directly connected with it. For instance consider the following chain of terms: *ukraine - kolesnikov - boris - storm - tehran - obama - visit*, extracted from G_{NA} constructed on 3rd June 2014 (see Evaluation section for further details). Assuming *storm* being the first query term, it is clearly seen that although its first-level neighbors refer to novel information on meteorological condition (to a tropical storm named Boris and to a dust storm in Tehran respectively) this relation is already lost on the second level. Terms *tehran - obama - visit* are more related to US-Iran politics, whereas *ukraine - kolesnikov - boris* to a pro-Russian unrest in Ukraine.

Thus firstly in the preparation step, the egocentric network [4] of depth 1 for the first query term (excluding the term itself) is extracted from novel associations graph to assure that the terms are semantically related. Therefore, the further processing is performed on a subgraph of novel associations graph consisting of the first query term's immediate neighbors (excluding the term itself).

Terms referring to the same topic or used in the same context should be relatively more densely connected with each other comparing to terms touching other subjects or employed in different situations. Hence, Newman's leading eigenvector method (based on modularity) for community structure detection [5] is then used to group together terms potentially referring to the same topic or subtopic. Next, the identified communities $(\mathbb{C})$ are sorted based to their sizes.

Central terms can allow to produce more query terminations (e.g. taking a term having a number of connections is better than taking a starting node of a long chain). What is more, such terms can help providing richer, more diversified suggestions by exploiting different parts of egocentric network (extracted for a given query term from the novel associations graph). This should allow to cover more semantic aspects. Hence, central terms can be considered good candidates for starting points of query terminations. On the other hand, the most novel terms co-occurrences can be located at any part of the graph. To account for both aforementioned aspects, the nodes' weighted *PageRank* values are calculated. Finally, the best nodes according to their weighted *PageRank* values are identified in each community $(\mathbb{C})$ and all paths of length 3 and 4 starting from those nodes are calculated to then generate query terminations[3].

Selection.

In the selection of query terminations step, k paths maximizing the average novelty score (weights of edges in the novel associations graph for a given path) are chosen as the query terminations. The paths are selected in the iterative manner, one at a time from each community starting from the biggest community. The process continues till k terminations are found or there's no more paths to select terminations from.

Formally, let $p = (t_1, \ldots, t_l)$ be a path consisting of $l \in \{3, 4\}$ terms and $\mathcal{P} = \{p_1, \ldots, p_r\}$ the set of all r paths identified in the given community. We define the total novelty score of a path p as:

$$s(p) = \sum_{i=1}^{l-1} w(t_i, t_{i+1}). \tag{5}$$

While generating query terminations, we maximize the following function:

$$\mathcal{T} = \arg\max_{p \in \mathcal{P}} \frac{s(p)}{l}, \tag{6}$$

in each of the detected communities.

3. EVALUATION

The approach to proactive, novel query recommendations presented in this paper has been preliminarily evaluated using a user study and an online survey. This section discusses the details on the experimental framework, experiment designs and obtained results.

Novel information can be easily found in news reports. Moreover, major search engines regularly crawl and index news agencies' sites. Thus, news domain was chosen for the initial performance comparison of the QAC method proposed in this article with existing industry solutions. Addi-

[3] This paper presents the work in progress. At this point it is limited to provide fixed-length suggestions. The evaluation is performed using suggestions of length 3 and 4. As the initial results are very promising this limitation is about to be removed giving more flexible solution.

tionally, even though the proposed query recommendations method is language independent the evaluation was performed for English. Its assessment in other languages as well as in other domains is left for future work.

The approach for QAC presented in this paper is grounded on the comparison of two, normative and sample, corpora. To gather data indispensable to create these corpora a crawler collecting news snippets from 4 major news agencies (CNN, REUTERS, BBC and The New York Times) was built. If the normative corpus was initially empty all terms associations would be treated as novel. Hence, it was initialized with 1 week data obtained from the aforementioned sources, specifically on data published between 27th May and 2nd June, 2014. Then, it was being incrementally expanded using newly crawled data. The sample corpus was being constructed on request using newly crawled data whose age didn't exceed 90 minutes. This time interval was found to be sufficiently long for news agencies to provide novel information and as so for the QAC system to produce some novel suggestions.

To evaluate the approach a user study was performed and a survey was carried out. In both cases the initial corpora described above (which were evolving with time passing) were used. The potential 100 test queries were being selected dynamically at random from the set of all terms constituting the current sample corpus reduced by all the terms occurring less than 5 times. In both the user study and the survey the participants were asked to choose a test query from a word cloud of these 100 randomly selected terms. Both experiments ended on 3rd June, 2014.

User Study.

In the user study 5 volunteers – experienced search engines users, were asked to judge top 5 recommendations generated by the method to 10 query terms of their choice. The volunteers were provided with a simple assessment interface in which the chosen query term and then in sequence query terminations generated for the given test query were presented. Each assessor was asked to decide whether a given query termination was novel and whether it could potentially support the proactive information search using self-developed 2-item novelty scale. No evaluation criteria were given. Additionally, the notion of novelty was not specifically explained allowing for multiple interpretations. The participants were encouraged to use a search engine of their choice to verify the suggestions before deciding on their quality.

Coverage of the proposed query suggestion method was then computed to assess its success rate. The method was considered successful if it was able to generate at least one novel and meaningful suggestion for a given query term (compare with the definition proposed by Bhatia et al. [8]).

Results indicated that the method was able to generate at least one novel and meaningful suggestion for all the tested queries (coverage = 100%). Nevertheless, the participants in their comments reported its relative deficiency in wording, it is in arranging terms forming query suggestion in the adequate order. This imperfection should not yet lead to prejudge the effectiveness of the method. The algorithm used to generate suggestions use neither natural language processing techniques nor linguistic resources or knowledge bases. Moreover, it operates on the very small dataset (consisting of news reports' snippets from 4 sources collected over a week). Thus, we believe that after additional fine-tuning this method will prove to be a great technique supporting proactive information search and retrieval. As so it will contribute to a reduction of cognitive load on users.

Survey.

To appraise the method in more real-life settings an online survey was conducted. The survey was available only for US and UK area to reduce the potentially negative effects of differences in English proficiency introduced by non-native English speakers on evaluation results. Search engines are used by very diversified groups of people. Thus no further restrictions (for example on respondents' age or highest education level achieved) were introduced. Forty participants were recruited via *CrowdFlower* crowdsourcing platform.

The subjects were presented top 3 recommendations generated by the method proposed in this paper as well as top 3 recommendations obtained from Google and Bing non-official services. They used the same simple assessment interface as the user study participants though they were presented query terminations from different sources. The source of the recommendations was not revealed to the assessors. The presentation order of the recommendations was also randomized to avoid potential ordering effects.

The participants were asked to evaluate the ability of the given suggestions to provide the users with new or emergent information using a self-developed 2-item novelty scale (see Table 2 and Table 3). As in the user study, no further instructions suggesting potential evaluation criteria or providing definitions of vital terms (like e.g. novel information) were given. Similarly, the respondents before drawing a conclusion were encouraged to check the search results generated by their favorite search engine for the given query recommendation.

The scale was found to be valid, with an excellent α score exceeding 90%. The results (see Table 2 and Table 3) indicate that the presented approach is already capable of competing with industry standards.

Table 2: Crowdsourcing evaluation results: Can the suggestion help in finding novel information?

	Google [%]	Bing [%]	Our method [%]
No	8.57	27.94	20.90
Yes	91.43	72.06	79.10

Table 3: Crowdsourcing evaluation results: Can the suggestion support exploratory or proactive information search (open-ended search)?

	Google [%]	Bing [%]	Our method [%]
No	4.29	27.94	10.45
Yes	95.71	72.06	89.55

Figure 1 and Table 4 show example results generated by the proposed QAC method and suggestions obtained from Google and Bing non-official services for the query term *storm* on 3rd June, 2014.

4. CONCLUSIONS AND FUTURE WORK

In this paper we have introduced a new approach for query auto-completion: proactive, novel query recommendations,

Table 4: Top 5 query suggestions for the term "storm" generated by our method and obtained from Google and Bing non-official services on June 3rd, 2014

Our method	Reference	Google	Bing
storm boris mexico forcing communities	tropical storm named *Boris* formed off Mexico's Pacific coast causing evacuation of seaside and mountain communities due to flooding and mudslides risk	**storm** chasers	**storm**.no
storm tehran deadly yesterday stunning	dust storm hit Tehran (Iran) on 2nd of June 2014 killing at least 5 people and causing stunning transformation in weather conditions	**storm** prediction center	**storm** no
storm salina cruz centered southeast	on 3rd of June 2014 *Boris* was centered about 125 miles southeast of Salina Cruz	**storm** bowling	**storm**
storm northward toward refinery churning	*Boris* storm was churning northward toward the largest Pemex (national oil company) refinery	**storm**	**storm**riders
storm boris mountain forcing communities	see explanation for the 1st suggestion	**storm**y castle	**storm** of swords

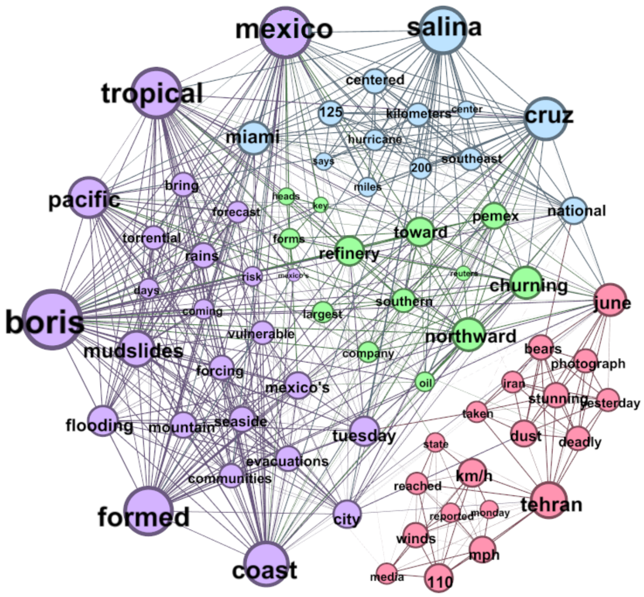

Figure 1: Ego-centric network extracted from the novel associations graph for the query term "storm" on 3rd June, 2014. Nodes represent terms and edges indicate terms co-occurrences. Nodes are sized according to their weighted PageRank values, and colored according to the communities they belong to. Edges' widths denote their novelty scores (*OR* — odds ratio values).

which aims at reducing the cognitive load on users by providing them with suggestions referring to novel, emergent information. This method should be considered as a complement to existent methods leveraging past knowledge (generally acquired by mining query logs) to generate suggestions.

This paper presents the work-in-progress. Simulations examining different methods of novelty detection and query terminations generation, as well as their impact on the performance of the proposed approach are necessary to be able to efficiently implement it in practical settings. All of the aforementioned aspects are currently being researched. However, in this paper we demonstrate that the identification of novel information and its provision to the users in the *push* manner (in the form of query suggestions) is achievable and can be effective even without using any linguistic resources or knowledge bases.

5. ACKNOWLEDGMENTS

This work has been supported by The French Ministry of Foreign Affairs through the Eiffel Excellence Scholarship and by Tekes, Tivit, and Algorithmic Data Analysis (Algodan) Center of Excellence of the Academy of Finland.

6. REFERENCES

[1] N. Carr. Real-Time Search. `http://www2.technologyreview.com/article/418536/tr10-real-time-search/`. MIT Technology Review, 2010.

[2] Y. Yang, K. Kim, B. Chung, J. Park. A Scalable Real-time Search Engine for Fast Retrieval of Social Media Content. In *Proceedings of the 2nd International Workshop on Ubiquitous Crowdsouring.* ACM, 2011.

[3] R. Hemayati, L. Dehkordi, W. Meng. mNIR: Diversifying Search Results based on a Mixture of Novelty, Intention and Relevance. In *Proceedings of WISE'12*. Springer-Verlag, 2012.

[4] J. Heer and D. Boyd, Vizster: Visualizing Online Social Networks. In *Proceedings of INFOVIS'05*. IEEE Computer Society, 2005.

[5] M. E. J. Newman Finding community structure in networks using the eigenvectors of matrices. Physical Review E, 74, 2006.

[6] J. Cornfield A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. Journal of the National Cancer Institute, 11(6), 1951.

[7] F. Mosteller. Association and Estimation in Contingency Tables. Journal of the American Statistical Association, 68(321) 1968.

[8] S. Bhatia, D. Majumdar, P. Mitra. Query Suggestions in the Absence of Query Logs. In *Proceedings of SIGIR'11*. ACM, 2011.

[9] M. Shokouhi, K. Radinsky. Time-Sensitive Query Auto-Completion. In *Proceedings of SIGIR'12*. ACM, 2012.

[10] P. Boldi, F. Bonchi, C. Castillo, D. Donato, S. Vigna. Query Suggestions Using Query-Flow Graphs. In *Proceedings of WSCD'09*. ACM, 2009.

[11] R. Baeza-Yates, Hurtado, Mendoza. Query Recommendation Using Query Logs in Search Engines. LNCS. Springer, 2004.

[12] I. Szpektor, A. Gionis, Y. Maarek. Improving recommendation for long-tail queries via templates. In *Proceedings of WWW'11*. ACM, 2011.

[13] G. A. Miller. WordNet: a lexical database for English. Communications of the ACM, 38(11) 1995.