Retrieval Techniques for Contextual Learning

Nino Weingart
ETH Zurich, Switzerland
Dept. of Computer Science
ninow@student.ethz.ch

Carsten Eickhoff
ETH Zurich, Switzerland
Dept. of Computer Science
ecarsten@inf.ethz.ch

ABSTRACT

Following constructivist models of contextual learning, knowledge acquisition goes beyond mere absorption of isolated facts, and, instead is enabled, stimulated and supported by related existing knowledge and experiences. In this paper, we discuss a range of query expansion and result list re-ranking techniques aiming to preserve contextual dependencies among retrieved documents and, thereby, enhancing the performance of learning-centric search engines.

Our empirical evaluation is based on a snapshot of Wikipedia and suggests significantly increased usability during an interactive user study.

Keywords

Query Expansion, Re-ranking, Wikipedia, Learning, Knowledge Acquisition

1. INTRODUCTION

The Internet connects vast numbers of knowledge resources that are assumed to contain the necessary information to answer most general questions occurring to common users [6, 10]. However, even given this seemingly inexhaustible well of information, the act of learning from this information is constituted by more than just memorizing facts. Constructivist theoreticians including Piaget and Vygotsky argue that the learning process necessarily depends on the context of existing knowledge upon which the newly encountered factoids are building. The stronger the contextualization of new knowledge, the more effortless and effective the learning is assumed to be.

Despite the wide acceptance and demonstrated success of constructivist methods in pedagogics, common retrieval models do not support any notion of *contextual learning*. Document relevance is largely judged in isolation and list-wide ranking considerations rarely go beyond diversification efforts. Consequently, state-of-the-art search engines cannot be considered ideal learning environments.

We believe that this shortcoming is manifested in threefold form: (1) Raw textual documents may not be the ideal retrieval unit. Due to high variance in length and an often non-uniform distribution of relevant factoids across documents, learning may be better supported by a finer granularity. (2) The user's learning intent by definition characterizes a degree of unfamiliarity with the desired information. To account for this fact, query formulation should be guided not just by user-specified terms but also by connections and dependencies dictated by the studied subject matter. (3) The probability ranking principle is based on point estimates of relevance. While this approach ensures maximum relevance at early result list ranks, it ignores important causal dependencies between documents, potentially resulting in a chaotic and didactically dissatisfying ordering of material.

While paragraph retrieval, query expansion and result list re-ranking are known techniques, this work aims to use and combine them to optimally support the goal of contextual learning in Web search. We believe that as such, this overview, along with the results of an empirical user study, will be of interest to the community.

The remainder of this paper is structured as follows: Section 2 gives a necessarily brief introduction into the concepts of contextual learning as well as domain expertise upon which this work builds. Section 3 formally describes three concrete techniques that address the previously enumerated shortcomings of state-of-the-art retrieval models. Section 4 puts these methods to use in a real-world learning-centric search scenario. Finally, Section 5 concludes with a brief discussion of our findings as well as an outlook on future directions in this domain.

2. RELATED WORK

This section gives a brief overview of related work dedicated to two directions. Beginning with a brief discussion of formal constructivist theories, we proceed to a more applied line of work dedicated to estimating user domain expertise during Web search.

Jean Piaget first proposed the theory of "Cognitive Development" that considers knowledge to be an actively constructed complex system of experience, stage of cognitive development, cultural background and personal history [9]. In other words, knowledge is derived from personal experience and ideas rather than an aggregation of loose facts and formulas. Building on Piaget's theories, Langley [8] studies order effects in incremental learning. The author claims that the order in which material is learned has a significant influence on the overall learning rate and absolute retention. Kuhlthau et al. [7] discuss the importance of mediators who enable learners to go beyond the current limits of their understanding. In this work, we try to deliver some of this mediating support by means of technological aids. Concretely, we aim to contextualise and order factoids in a more

Search as Learning (SAL), July 21, 2016, Pisa, Italy

supportive way than dictated by state-of-the-art models.

Over the past years, the study of user's existing domain-specific knowledge has led to a wide number of innovations in user understanding. White et al. [13] show that, within their area of expertise, domain experts search differently from non-experts. They are found to use a more diverse vocabulary of query terms and generally demonstrate a better understanding of the desired results to be retrieved, resulting in improved query formulation and result inspection performance as compared to laypeople. Eickhoff et al. [4] observe that the search behavior of a user changes over time. This is assumed to occur as a consequence of having learned while searching and therefore having acquired increased domain expertise. To promote fast learning, and thus changing search behaviour early on, the authors suggest identifying key terms that help to improve their vocabulary. In a follow-up study [3], this hypothesis is further evidenced with the aid of eye-tracking hardware, measuring term-level knowledge acquisition as users search the Web with the goal of learning about a previously unfamiliar topic. While, in this work, we do not explicitly model user domain expertise, the existing work in this direction serves as further evidence of the importance of contextualising search for learning. Concretely, we propose a result re-ranking mechanism that aims to support contextual learning including order and contextual dependencies [5]. Further, we propose a query expansion mechanism to guide expert and non-expert users to better search results as elaborated in studies by White et al. [13] and Eickhoff et al. [4] that expect an accelerated gain in domain expertise.

3. METHODOLOGY

In this section, we formally describe three techniques supporting contextual learning during search. Beginning with a paragraph retrieval model, we move on to pseudo-relevance feedback-based query-expansion as well as a method for dependency-based result list re-ranking.

3.1 Paragraph Retrieval Model

At the core of our method, we rely on a standard tf-idf model scoring documents d in response to a query q according to the frequency at which query term t occurs in d. While this score $tf_{t,d}$, is higher for documents that contain more query terms it has the drawback of treating all terms as equally important. To apply a non-uniform term weighting, defined by the specificity of a term throughout the collection, we further introduce the notion of t's *document frequency* df_t as the number of unique documents containing t. With both components in place, our retrieval model score s is given by

$$s(q,d) = \sum_{t \in q} \mathrm{tf}_{t,d} \times log \frac{N}{\mathrm{df}_t} \times w_t \tag{1}$$

where N denotes the total number of documents in the collection and w_t is an additional term weight that is uniformly set to 1 for all original query terms. In the following section we will discuss a query expansion scheme that may add new terms at a $w_t \neq 1$.

In this work, we attempt to better contextualise search results by considering documents of varying granularity. Depending on the concrete model, d can either be a complete document or a single paragraph extracted from a longer text. We assume that there are reliable techniques for breaking up documents into paragraphs. Section 4 will introduce the concrete paragraph extraction scheme that was applied in our experimental setup.

3.2 Query Expansion

Formulating effective queries has been shown to be a hard task that requires intimate familiarity with the subject domain as well as the underlying document collection [2]. In learning-centric search settings on the Web, neither of these prerequisites can be assumed to hold.

To address this issue, we introduce a *pseudo relevance feedback (PRF)* step. Instead of following the established approach by Rocchio [11], this work, instead relies on the topology of the link graph. This choice is motivated by the intuition of hyperlink referral expressing contextual dependency. *I.e.*, if the same document is referred to multiple times by a number of high-ranking documents, it is likely that the document represents a required resource to understanding the subject matter of the retrieved documents even if direct keyword matching on the original query does not discover this link.

We formally capture this intuition by collecting the set of top-k most highly scoring documents $D_{q,k}$ according to the original query q. For each document in this set, we follow outgoing hyperlinks and collect the bag of words T_{new} of all terms appearing in the titles of linked-to documents. Intuitively, T_{new} represents a description of the required reading for the original selection of highly relevant documents. To account for the relative importance of the newly added terms, we normalize the contribution of each term by $|T_{new}|$, the size of the bag of words. In this way, terms that occur in multiple titles or titles that are linked to frequently are weighted more prominently than singleton occurrences.

At this point, due to our normalization scheme, the sum of all newly added terms amounts to the same cumulative weight as a single original query term. In order to control the relative importance of original and newly added terms, the parameter w_{add} determines the "number" of virtual terms to be added.

$$w_t = \begin{cases} 1, & \text{for } t \in q \\ \frac{w_{add} \times count(t, T_{new})}{|T_{new}|}, & \text{otherwise} \end{cases} \tag{2}$$

Finally, the expanded query q' is given by the original query q as well as the linked title terms T_{new}, which are added with their respective importance weights w_t.

$$q' = q \cup T_{new} \tag{3}$$

3.3 Dependency based Re-Ranking

Web search-based learning can result in jumping back and forth between documents. This may happen because a document covers material that possibly requires previous knowledge from another source indicated by a reference. This induces a dependency structure over the documents. As motivated in Section 2, the order of material can have significant implications on learning rate and knowledge retention. As consequence, we would like to present documents in an ordering that respects this referral structure.

We again consider $D_{q,k}$, the top-k retrieved documents and compute the number of times each document is referred to from within $D_{q,k}$. Let us call this quantity the document's link count $c(d)$.

Table 1: Per-query and per-system survey results.

T	Q20	Q22	Q33	Q36	Q38	Q40	Q45	Q75	Q71	Q99	Q107	$\frac{1}{11}\sum$
A	65.94%	40.32%	47.76%	56.17%	61.20%	52.76%	70.34%	47.76%	55.19%	63.06%	68.24%	57.16%
AE	60.73%	45.45%	35.25%	47.73%	48.83%	55.03%	66.96%	41.51%	43.24%	59.09%	72.78%	52.42%
AR	64.18%	52.22%	46.62%	64.83%	57.21%	65.29%	68.06%	40.90%	38.13%	57.45%	71.05%	56.90%
AER	64.74%	59.58%	42.65%	62.01%	56.11%	64.10%	70.34%	46.57%	48.30%	63.13%	73.33%	59.17%
P	57.42%	55.65%	44.86%	61.44%	73.21%	64.14%	68.10%	51.73%	61.56%	63.15%	61.39%	60.24%
PE	47.11%	45.99%	39.78%	58.60%	56.70%	68.67%	68.72%	52.89%	62.02%	64.27%	65.43%	57.29%
PR	59.65%	58.52%	44.91%	71.62%	70.96%	69.77%	70.96%	59.15%	56.36%	65.40%	66.44%	63.07%
PER	53.97%	49.49%	42.07%	68.80%	52.11%	62.39%	62.96%	44.35%	52.33%	61.99%	64.26%	55.88%
$\frac{1}{8}\sum$	59.22%	50.90%	42.99%	61.40%	59.54%	62.77%	68.30%	48.11%	52.14%	62.19%	67.86%	

$$l(d) = \begin{cases} \beta c(d), & \text{for } d \in D_{q,k} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where β is a scaling weight to boost the resulting link score $l(d)$ into the regime of $s(q,d)$. Our final dependency-aware retrieval model score $s'(q,d)$ is given by a weighted mixture of the original score s as well as $l(d)$, the document's importance in the context of $D_{q,k}$.

$$s'(q,d) = \alpha s(q,d) + (1-\alpha) l(d) \quad (5)$$

4. EXPERIMENTS

Our empirical investigation of the practical usefulness of the presented methods is based on a recent snapshot of Wikipedia. The platform represents a popular knowledge resource and is consulted at high frequency every day. We provide a basic retrieval system implemented in the Apache Lucene framework[1] and index either full articles or paragraphs as the atomic retrieval unit. Due to Wikipedia's article structure, paragraph splitting is a straight-forward process guided by the original article's sections (identified by "==" and "===" delimiters). To ensure realistic and informative document titles, paragraphs concatenate their parent article's title with their own section heading.

We study the following $2^3 = 8$ combinations of experimental conditions:

- Document vs. paragraph retrieval (2)
- With or without query expansion (2)
- With or without re-ranking (2)

To evaluate these conditions, we select 10% (11 out of 107) of the INEX 2010 *Ad Hoc* queries [1] to cover a diverse range of topics and query lengths. For each topic, we generate a survey for every query consisting of 5 multiple choice questions with six possible answers. The number of correct answers varies between one to four out of the six choices. We now displayed the topic narrative as well as the results retrieved by one of the experimental conditions to *Amazon Mechanical Turk* workers and subsequently asked them to complete the corresponding survey, answering questions to the best of their knowledge.

[1]https://lucene.apache.org/

Table 2: Relative effect of paragraph retrieval.

Reference System	$\text{Score}_{Article}$	$\text{Score}_{Parameter}$	Δ
A	57.16%	60.24%	+5.39%
AE	52.42%	57.29%	+9.30%
AR	56.90%	63.07%	+10.83%
AER	59.17%	55.88%	-5.56%
Mean	56.41%	59.12%	+4.99%
Std. Deviation	2.47%	2.77%	6.40%

Each question is scored according to the number of correct answers divided by the total number of choices, where "correct" refers to either the selection of a correct answer choice or the leaving blank of a wrong option. To further penalize random guessing, we subtract points for wrong answers, bounding the score per question in $[0,1]$.

$$score = \frac{max(\#correct - \#wrong, 0)}{\#correct + \#wrong} \quad (6)$$

With this scoring scheme in place, workers, topics and experimental conditions can be evaluated by averaging across all scores of the respective selection.

4.1 Results

For each of the 8 experimental conditions and 11 topics we collect answers from 5 individual workers, leading to a total of 440 experiment submissions.

Table 1 shows the scores per experimental condition. Topic numbers are denoted in columns. The experimental conditions are encoded according to the respective components used. "A" and "P" refer to full article vs. paragraph retrieval. "E" indicates the use of query expansion and "R" the dependency based re-ranking. The right-most column details each system's mean score across topics. Similarly, the bottom-most row contains mean scores per topic across all systems, capturing the difficulty of each topic and the corresponding questions.

The correctness scores range from 35.25% to 73.33% with the overall mean score being 57.77% (median 59.37%). At a glance, we note a considerable variance in the difficulty of individual topics, while the performance of the compared

Table 3: Relative effect of query expansion and result list re-ranking on answer correctness scores.

Reference System	Score_{old}	Score_{+E}	Δ	Reference System	Score_{old}	Score_{+R}	Δ
A	57.16%	52.42%	-8.30%	A	57.16%	56.90%	-0.45%
P	60.24%	57.29%	-4.90%	P	60.24%	63.07%	+4.69%
Baseline Mean	58.70%	54.85%	-6.60%	Baseline Mean	58.70%	59.98%	+2.12%
AR	56.90%	59.17%	+3.98%	AE	52.42%	59.17%	+12.88%
PR	63.07%	55.88%	-11.39%	PE	57.29%	55.88%	-2.45%
Re-Ranking Mean	59.98%	57.53%	-3.70%	Expansion Mean	36.91%	26.72%	+5.22%
Total Mean	59.34%	56.19%	-5.15%	Total Mean	56.78%	58.76%	+3.67%
Std. Deviation	2.52%	2.47%	5.75%	Std. Deviation	2.80%	2.76%	5.92%

systems is more closely tied. For greater ease of inspection, in the following, we provide a number of detailed views extracted from the overall data.

Let us begin by evaluating the effect of document granularity. Table 2 shows the performance difference observed when switching from retrieving full documents to paragraphs in otherwise identical experimental conditions. The scores show that for most experiment conditions, performance scores are significantly greater when retrieving paragraphs instead of full articles. Only the re-ranked and expanded condition (AER vs. PER) performs better when retrieving articles. On average, switching to paragraph retrieval introduced a 4.99% increase in scores. We suspect that the reason for this tendency may lie in the different document lengths of articles and paragraphs. Highly ranked paragraphs provide a high density of relevant information whereas full articles can contain lengthy stretches of unrelated content.

Table 3 highlights the effect of adding query expansion or result list re-ranking to a reference system. The first column indicates the reference system, to which we add either query expansion (E) or result list re-ranking (R), while keeping all other conditions stable. We can observe that dependency-based re-ranking has a mild positive effect on the users' correctness scores while query expansion, in the vast majority of investigated conditions, shows a negative effect. We suggest that the query expansion mechanism as implemented, cannot effectively address the wide variety of query subjects and lengths. It can conceivably benefit from an interactive implementation, where a user can take influence on the term selection and determine the impact upon its level of expertise as elaborated by Vakkari [12]. This may produce better single search results and increase the contextual learning effect by individually enhancing the mechanism for a given query alongside the benefit from taking action into the search process as a whole. The results of dependency based re-ranking tend to produce more precise top-ranked search results. For contextual learning effects we would need to study the overall effect over time instead of a point estimation of knowledge acquisition. For both of these results, it should be noted that we observe a considerable amount of variance between runs, suggesting that a more large-scale investigation may be in order before conclusive insight can be gained.

5. CONCLUSION

In this paper, we present initial results of an ongoing investigation into the suitability of established retrieval techniques as well as variants thereof, for the task of contextual learning in Web search. Inspired by constructivist theories of learning and knowledge formation, we propose a paragraph retrieval model, a document title based query expansion scheme as well as a result list re-ranking method that aims to preserve order dependencies in the material.

We conducted a learning-centric user study on the basis of the Wikipedia corpus during which participants used varying combinations of the above components to help their search sessions. The experiment showed a strong positive effect of using paragraphs instead of full documents as retrieval units. While the query expansion approach turned out to result in an overall negative effect, dependency based re-ranking resulted in an increased performance score on average. While this study is limited in both the size of the user base as well as the diversity of information needs, it shows promising potential and highlights the importance of explicitly accounting for contextual learning during the retrieval process.

There are several exciting directions of future inquiry. To ensure comparability between limited-scale results, the present study relies on fixed queries and only incorporates real users as result list consumers. While this paradigm showed good results, it would be interesting to study the proposed techniques in a truly interactive search setting in which the users themselves formulate their queries. In such a setting, one can imagine a wide range of interesting controls that enable the user to specify the exact amount (*e.g.*, in terms of number of pages or minutes worth of reading) of material to retrieve as well as its topical focus. Further, a user could interactively adjust a wide range of parameters for the search engine presented in Section 3 or interactively select additional query terms to obtain better search results. According to Piaget's "Cognitive Development" theory, the contextual learning effect can increases with the possibility to actively participate in the learning step, *e.g.*, the searching step for learning-based search engines.

Additionally, the present study focuses on point estimates of factual knowledge acquisition. In the future it would be interesting to conduct more longitudinal investigations of learning rates and knowledge retention in the true constructivist spirit.

Finally, the task of contextual learning in Web search is an exciting environment for user modelling and personalization efforts in which notions such as domain expertise or preferred reading levels will play a key role.

6. REFERENCES

[1] P. Arvola, S. Geva, J. Kamps, R. Schenkel, A. Trotman, and J. Vainio. Overview of the INEX 2010 ad hoc track. In *Comparative Evaluation of Focused Retrieval*, pages 1–32. Springer, 2010.

[2] N. J. Belkin, R. N. Oddy, and H. M. Brooks. ASK for information retrieval: Part I. Background and theory. *Journal of documentation*, 38(2):61–71, 1982.

[3] C. Eickhoff, S. Dungs, and V. Tran. An eye-tracking study of query reformulation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 13–22. ACM, 2015.

[4] C. Eickhoff, J. Teevan, R. White, and S. Dumais. Lessons from the journey: A query log analysis of within-session learning. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 223–232. ACM, 2014.

[5] E. B. Johnson. *Contextual teaching and learning: What it is and why it's here to stay*. Corwin Press, 2002.

[6] J. Kräenbring, T. M. Penza, J. Gutmann, S. Muehlich, O. Zolk, L. Wojnowski, R. Maas, S. Engelhardt, and A. Sarikas. Accuracy and completeness of drug information in wikipedia: a comparison with standard textbooks of pharmacology. *PloS one*, 9(9):e106930, 2014.

[7] C. C. Kuhlthau. Seeking meaning. *Norwood, NJ: Ablex*, 1993.

[8] P. Langley. Order effects in incremental learning. *Learning in humans and machines: Towards an interdisciplinary learning science. Pergamon*, 136:137, 1995.

[9] J. Piaget. *Piaget's theory*. Springer, 1976.

[10] N. J. Reavley, A. J. Mackinnon, A. J. Morgan, M. Alvarez-Jimenez, S. E. Hetrick, E. Killackey, B. Nelson, R. Purcell, M. B. Yap, and A. F. Jorm. Quality of information sources about mental disorders: a comparison of wikipedia with centrally controlled web and printed sources. *Psychological medicine*, 42(08):1753–1762, 2012.

[11] J. J. Rocchio. Relevance feedback in information retrieval. 1971.

[12] P. Vakkari. Subject knowledge, source of terms, and term selection in query expansion: An analytical study. In *European Conference on Information Retrieval*, pages 110–123. Springer, 2002.

[13] R. W. White, S. T. Dumais, and J. Teevan. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 132–141. ACM, 2009.