Improved Cascade for Search Mission Detection

Matthias Hagen, Jakob Gomoll, and Benno Stein

Bauhaus-Universität Weimar
<first name>.<last name>@uni-weimar.de

Abstract Search sessions are formed by consecutive queries that users submit to a search engine when addressing the same information need. However, sessions of consecutive queries may not comprise all queries that belong together, as users may interleave different search tasks in a multitasking manner or as some search sessions are just smaller parts of some larger search mission.
Our contributions in this paper are threefold. First, we present a new algorithm for session detection based on the state-of-the-art cascading method. We improve the single steps of the cascading method and, for the first time, we exploit Linked Open Data information for session detection. With the second main contribution we show that our new algorithm is also able to detect multitasking situations as well as search missions. Finally, our third contribution is the development of a new, publicly available corpus of more than 8800 queries manually labeled with search session and mission information. This new corpus exceeds the size of the largest previously available mission detection corpus by an order of magnitude.

1 Introduction

Query session detection is the problem of identifying the series of consecutive queries a user submits for the same information need. Studying such sessions is interesting for getting insights on how users search and in which regard they encounter problems. However, considering only consecutive queries will miss important connections: typical patterns include interleaving sessions resulting from a multitasking search behavior as well as hierarchies of different search goals and so-called "missions" [7, 10, 11]. A user may for instance shortly interrupt a longer search session for checking the weather forecast, or she may follow subordinate search goals spanning several days which finally form a larger mission (e.g., planning the next vacation).

In this paper we address both problems, search session detection and multitasking/mission identification. Therefore, we improve the recent stepwise approach of the cascading method [4] with a new step that employs a Linked Open Data analysis for session detection. We then suggest a two phase application of the improved cascade to detect multitasking behavior and search missions: in a first phase the sessions are detected in the traditional way, and in a second phase the detected sessions are checked for further connections. We evaluate the improved cascade on a new large scale corpus of 8800 queries annotated with search session and mission information.

2 Framework and Related Work

For each query q a search engine log contains the query string, a user ID and a time stamp denoting when q was submitted. Session detection is typically modeled as the

problem of deciding for each consecutive pair q, q' of queries from one user whether the session s that contains q continues with q' or whether q' starts a new session.

Early methods combined queries into sessions whenever the elapsed time between their submission was below a certain threshold [5]. Such sessions were typically not meant to reflect the "modern" understanding of queries with the same information need, but to obtain those queries that users submitted together. Hence, it is not too surprising that the time-based methods achieve only about 70% accuracy in recent, real-world session detection tasks [7]. Current state-of-the-art methods combine additional features to measure the syntactic or semantic similarity of query pairs. One of the best performing approaches with respect to the detected sessions' accuracy is the cascading method [4]. It sensibly involves different features in different steps considering the increasing computational costs. This way, the cascading method is much faster than other methods that combine all features simultaneously without hurting the session accuracy.

However, session detection as applied by the cascading method can neither cope with users who interleave different sessions in a multitasking manner nor with hierarchies of search goals and missions. In both cases the cascading method would just identify the small (interleaved) pieces of longer sessions and missions without recognizing their connections. Only two approaches for multitasking and mission detection are published yet [7, 8]. Both apply the entire feature set at once albeit the cascading method showed that deferring the exploitation of computationally costly features is beneficial [4]. Our proposed new approach to multitasking/mission detection adopts the cascading method's strategy and improves the method in two respects: by tailoring its steps and by adding a new step for the semantic similarity check of two queries. The new step is based on an analysis of Linked Open Data (LOD) connections, a technique previously used for the identification of typical query reformulation patterns [6].

3 A New Detection Method

The version of the original cascading method that we use in this paper consists of three steps with increased feature (computation) cost from step to step [4]. In the very efficient first step, query pairs that are repetitions, generalizations, or specializations are put in the same session. The second step involves the geometric method [2] that combines both lexical similarities and the time elapsed between two queries. Whenever the time is below a predefined threshold and the lexical similarity is high, the two consecutive queries are assigned to the same session. However, when only a short period of time passed between two consecutive queries and the lexical similarity is low, the second step will miss semantically related queries. To address this problem the cascade involves an ESA step (Explicit Semantic Analysis) [1] to identify semantically related query pairs.

3.1 Improving the Second Step: The Geometric Method

Based on a 25% training set sample of our newly developed query corpus (cf. Section 4), we evaluated the original three step cascade and made the following observations.

In its second step, the original cascade applies the geometric method by Gayo-Avello [2]. This method computes a time feature $f_{\text{time}} = \max\{0, 1 - \frac{t'-t}{24h}\}$ for the

submission times t, t' of a pair q, q' of consecutive queries as well as the cosine similarity $f_{\cos}$ between the character 3-/4-/5-grams of the query q' and the session s whose current last query is q. The geometric method votes for a session continuation iff $\sqrt{(f_{\text{time}})^2 + (f_{\cos})^2} \geq 1$. Whenever $f_{\cos} < 0.4$ and $f_{\text{time}} > 0.8$, the original cascade does not trust the geometric method's decision but invokes the third step (ESA). The rationale is that these are query pairs relatively close in time that could still be semantically related and thus belong to the same session—despite their low n-gram similarity.

In our improved second step we change the feature computation, the session continuation condition, and the above described "trust" range for invoking the ESA step. As for f_{time} we identified on our training set that a maximum time gap of 18 hours suffices, i.e., $f_{\text{time}} = \max\{0, 1 - \frac{t'-t}{18h}\}$. For $f_{\cos}$ we use only character 3- and 4-grams which saves time while not impairing the session accuracy on our training set. Another time saving aspect is that our improved Step 2 votes for a session continuation whenever $f_{\text{time}} + f_{\cos} \geq 1$ (three arithmetic operations less per query pair). Finally, we invoke Step 3 (ESA) of the original cascade whenever $f_{\cos} < 0.12$ and $f_{\text{time}} > 0.93$ (values identified by a grid search with steps of 0.01) and otherwise trust the Step 2 decision.

3.2 A New Fourth Step for the Cascading Method

In addition to the improvements on the original second step described above, we equip the cascading method with a new fourth step. Within this new step, the semantic similarity of a pair q, q' of consecutive queries is analyzed via the Linked Open Data (LOD) graph of DBpedia.[1] Hollink et al. used Linked Open Data for detecting typical reformulation patterns in an image search query log and suggested to test the potential for session detection [6]—an idea that we now pick up. DBpedia contains RDF triples representing entities and relations between entities mined from Wikipedia. Viewing entities as nodes and relationships as edges, the data defines a graph. Finding a path in this graph from one entity to another is a means of establishing the semantic similarity between these two entities. Typically, the more paths between two entities exist and the shorter these paths are, the more similar are the two entities. There is for instance a single step path from `david beckham` to `victoria beckham` since they are married.

To efficiently find paths in the DBpedia graph, an index stores the entities and the relationship RDF triples. For each entity e there is a postlist that contains all entities occurring in an RDF triple with e. A postlist entry p also has an idf-like weight of $\log\left(\frac{pl}{pl_p}\right)$, where pl is the total number of postlists and pl_p is the number of postlists that contain p. The idea of these weights is to discount paths containing very frequent relationships (e.g., all people entities have a relationship to their respective home countries such that a path including a country is a rather weak indicator of semantic relatedness).

The new fourth step works as follows. First, the main DBpedia entities in the queries q and q' are identified via query segmentation [3]. Thereby, the main entity is the query segment with highest web frequency that can be mapped to a DBpedia entity. For the mapping, we employ a dictionary that unifies different "names" of an entity to the same generic entity in the LOD graph (e.g., `david r. beckham` is unified to `david beckham`). The LOD step then identifies all two step paths (i.e., at most one

[1] http://dbpedia.org/

intermediate entity) from the main entity e in q to the main entity e' in q'. This can be accomplished efficiently by merging the indexed postlists of e and e'. The queries q, q' are assigned to the same search session iff the summed weight of all found paths (i.e., the summed weight of the entities in the merged postlist) is larger than 2.9 (value determined on our 25% training set). Otherwise, q' is viewed as the start of a new session.

3.3 The Cascade for Multitasking / Mission Detection

The improved cascade is not only able to detect sessions. When applied in a two-phase process, the improved cascade can also be used to detect multitasking and larger missions. The first phase runs the cascade at query level in order to detect simple sessions. The second phase then checks two non-consecutive sessions by running the last query of the first session and the first query of the second session through the cascade. Both sessions are assigned the same mission (this also includes multitasking) whenever the two queries would be assigned to the same session by the cascade in the second phase.

4 Experimental Evaluation

The accuracy of detected sessions or missions is usually evaluated against query logs manually labeled with session/mission information—the ground truth corpus.

4.1 A New Large Scale Session/Mission Corpus

The only publicly available session detection corpus is published by Gayo-Avello [2], consisting of 11 484 queries from 215 users sampled from the 2006 AOL query log [9] with respect to the representativeness of typical querying behavior (ratio of repeated queries, click through rate, etc.). A single human annotator manually subdivided the sample into 4040 sessions with an average of 2.70 queries per session. The main drawbacks of Gayo-Avello's corpus are (1) that query submission times and clicks have to be reconstructed from the original AOL log as they are not included, (2) that several users contained in the sample submitted so few queries that session detection for them makes no sense at all, (3) that sometimes the ordering of the queries in the sample does not reflect the original ordering in the AOL log, (4) that queries from some users are left out with no reason, and finally, (5) that there are a lot of annotation errors regarding sessions splits that belong together or queries in a session that do not belong together.

The only publicly available mission detection corpus is published by Lucchese et al. [8], consisting of 1424 queries from 13 users also sampled from the AOL log. However, as is the case with the Gayo-Avello sample, not all queries from the users are contained in the log but only those periods of querying where no gap between two subsequent queries is longer than 26 minutes. This means that about 97% of the original queries from the 13 users in the AOL log are discarded from the sample. As a result, the Lucchese sample does not reflect typical querying behavior with respect to repeated queries etc., and it is rather small compared to Gayo-Avello's sample.

The described drawbacks render the available corpora not fully representative of real user behavior which causes doubts with respect to the reliability of experimental evaluations based on these corpora. Hence, to reliably evaluate our method, we do not

use the existing corpora but create a new one. To still ensure some level of comparability with previous studies, we choose the large Gayo-Avello sample as our basis. From the AOL log we extracted all queries of the 215 users contained in the Gayo-Avello sample. We removed all queries that are empty or just a URL (probably submitted by users mixing up the search field with the address bar) and all queries from the 88 users that submitted less than 4 queries (too few queries for reasonable sessions). The final query sample contains 8840 queries from 127 users. Two human annotators divided this sample into 2280 sessions and 1397 missions. Cases where the annotators did not agree initially were discussed to reach a consensus. This new Webis Search Mission Corpus 2012 (Webis-SMC-12) will be made freely available.[2]

On average, a user in our corpus conducts 11 missions with 6.33 queries each. A mission on average contains 1.63 sessions. Missions and sessions are interrupted by a different session or mission and later picked up again in 1009 cases by 80 users. About 76% of all the missions are finished within one day; the longest mission lasts 60 days.

4.2 Performance of the Improved Cascading Method

To evaluate the improved cascade, we use as the test set the 75% of our new Webis-SMC-12 corpus that were not used as the training set for the cascade's steps in Section 3.

Session detection. We compare our improved cascade to the original three step variant. The precision and recall of the session break and continuation decisions are measured via the F-Measure. By setting $\beta = 1.5$ we render wrong session continuations as a bigger problem compared to wrong session breaks. The original three step cascade achieves an F-Measure of 0.875. Our improved cascade achieves an F-Measure of 0.890 (statistically significant difference according to a t-test with confidence level $p = 0.05$). However, the main reason for the improvement is not the new LOD step but the improved variant of Step 2. The LOD step finds only two continuations that the three previous steps do not detect. The main reason is the characteristic of the queries in our sample. An error analysis shows that most query pairs on which the LOD step is invoked, are really hard pairs with no LOD entity at all. Also note that with about 1 ms per query pair on a standard quad-core PC, Step 4 currently is much slower than Step 2 or 3 with less than 0.2 ms, not to mention Step 1 with less than 0.01 ms.

Thus, to show the effectiveness of the LOD step we also compare it to the ESA-step on a manually developed sample of 100 pairs of person and place names that are semantically related. On this set, the LOD analysis identifies 77 of the semantically related pairs while the ESA-step can only identify 59. Thus, despite its rather "weak" performance on the query set, we suggest to keep the LOD analysis in the cascade. Potential further improvements of its time efficiency and effectiveness could be pruning the LOD graph and indexing complete Wikipedia articles (cf. Section 5).

Mission detection. As for the search mission evaluation, we run the second phase of the mission detection cascade on the sessions annotated in the the test set (75% of the Webis-SMC-12 corpus). The cascade correctly identifies 556 of the 726 mission

[2] http://www.webis.de/research/corpora

continuations in the test set missing only 170 continuations. This clearly shows the potential of the improved cascade for multitasking / mission detection. However, there also is an error rate of 97 sessions that are wrongly assigned to be a continuation by the cascade. Future work should thus focus on analyzing the errors made in order to further improve the cascade's mission accuracy.

5 Conclusion and Outlook

We have presented an improved cascading method for session detection that can also be applied to multitasking and mission detection. The improvements relate to the original cascade's second step and a new Step 4 that checks the semantic similarity of two queries based on a Linked Open Data (LOD) analysis. Just like with the original cascade, time-consuming features are applied only when the cheaper features failed to provide a reliable session detection.

As for the evaluation, we have developed the new, publicly available Webis-SMC-12 corpus of 8800 queries annotated with session and mission information. On this corpus, our improved cascade outperforms the original method with respect to the detected sessions' accuracy while being comparably efficient. Our experiments have also demonstrated the potential of the improved cascade for the detection of multitasking behavior at user side and for the identification of smaller parts of larger search missions.

Future research should address a more thorough error analysis resulting in a fine-tuning of the improved cascading method. One idea could be to speed up the LOD step by pruning the underlying graph in a preprocessing. Another interesting point could be a fifth step that has an index of the complete Wikipedia to check the pages of the LOD entities for semantic similarity instead of just the connections in the LOD graph.

Bibliography

[1] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proc. of IJCAI 2007*, pp. 1606–1611.
[2] D. Gayo-Avello. A survey on session detection methods in query logs and a proposal for future evaluation. *Information Sciences*, 179(12):1822–1843, 2009.
[3] M. Hagen, M. Potthast, B. Stein, and C. Bräutigam. Query segmentation revisited. In *Proc. of WWW 2011*, pp. 97–106.
[4] M. Hagen, B. Stein, and T. Rüb. Query session detection as a cascade. In *Proc. of CIKM 2011*, pp. 147–152.
[5] D. He and A. Göker. Detecting session boundaries from web user logs. In *Proc. of BCS-IRSG Colloquium 2000*, pp. 57–66.
[6] V. Hollink, T. Tsikrika, and A. P. de Vries. Semantic search log analysis: a method and a study on professional image search. *JASIST*, 62(4):691–713, 2011.
[7] R. Jones and K. L. Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proc. of CIKM 2008*, pp. 699–708.
[8] C. Lucchese, S. Orlando, R. Perego, F. Silvestri, and G. Tolomei. Identifying task-based sessions in search engine query logs. In *Proc. of WSDM 2011*, pp. 277–286.
[9] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *Proc. of Infoscale 2006*, paper 1.
[10] A. Spink, H. C. Özmutlu, and S. Özmutlu. Multitasking information seeking and searching processes. *JASIST*, 53(8):639–652, 2002.
[11] A. Spink, M. Park, B. J. Jansen, and J. O. Pedersen. Multitasking during web search sessions. *Information Processing & Management*, 42(1):264–275, 2006.