# Overview of the TREC 2012 Session Track

Evangelos Kanoulas* Ben Carterette† Mark Hall‡ Paul Clough§ Mark Sanderson¶

## 1 Introduction

The TREC Session track ran for the third time in 2012. The track has the primary goal of providing test collections and evaluation measures for studying information retrieval over user *sessions* rather than one-time queries. These test collections are meant to be portable, reusable, statistically powerful, and open to anyone that wishes to work on the problem of retrieval over sessions.

The experimental design of the track was similar to that of the second year [4]:

- sessions were real user sessions with a search engine that include queries, retrieved results, clicks, and dwell times;
- retrieval tasks were designed to study the effect of using increasing amounts of user data on retrieval effectiveness for the $m$th query in a session.

There were a few changes compared to the second year of the track, mostly regarding the topic construction and the relevance assessments:

1. topics were constructed by the track co-ordinators; past TREC topics (mainly from QA2007 and MQ2009) tracks were used as in the second year; however, for each past topic 2-4 new topics were generated with certain characteristics regarding the type of the task;
2. relevance was defined at the level of the overall topic instead of subtopics.

This overview is organized as follows: in Section 2 we describe the tasks participants were to perform. In Section 3 we describe the corpus, topics, and sessions that comprise the test collection. Section 4 gives some information about submitted runs. In Section 5 we describe relevance judging and evaluation measures, and Sections 6 present evaluation results and analysis.

*Google, Zurich, Switzerland
†Department of Computer & Information Sciences, University of Delaware, Newark, DE, USA
‡Information School, University of Sheffield, Sheffield, UK
§Information School, University of Sheffield, Sheffield, UK
¶Department of Computer Science & Information Technology, RMIT University, Melbourne, Australia

## 2 Evaluation Tasks

We use the word "session" to mean a sequence of reformulations along with any user interaction with the retrieved results in service of satisfying an information need. The primary goal for participants of the 2012 track was to provide the best possible results for the $m$th query in a session given data from the session leading up to it.

NIST provided a set of 98 sessions of varying length (described in more detail in Section 3). Each session consists of:

- the current query $q_m$;
- the query session prior to the current query:
    1. the set of past queries in the session, $q_1$, $q_2$, ..., $q_{m-1}$;
    2. the ranked list of URLs for each past query;
    3. the set of clicked URLs/snippets and the time spent by the user reading the corresponding to each clicked url webpage.

Participants then ran their retrieval systems over only the current query under each of the following four conditions separately:

**RL1** ignoring the session prior to this query

**RL2** considering only the item (1) above, i.e. the queries prior to the current

**RL3** considering only the items (1) and (2) above, i.e. the queries prior to the current along with the ranked lists of URLs and the corresponding web pages

**RL4** considering all the items (1), (2) and (3) above, i.e the queries prior to the current, the ranked lists of URLs and the corresponding web pages and the clicked URLs and the time spent on the corresponding web pages

Comparing the retrieval effectiveness in (RL1) with the retrieval effectiveness in (RL2)–(RL4) one can evaluate whether a retrieval system can use increasing amounts of information prior to a query to improve effectiveness for that query.

## 3 Test Collection

Our test collection consists of a corpus, a set of topics, and relevance judgments (described in the next section). But unlike most test collections, ours also includes a set of *sessions* of user interactions (including query reformulations). A single topic can have more than one session associated with it, since two different sessions could go about satisfying the same information need in very different ways and with different degrees of success.

### 3.1 Corpus

The track used the ClueWeb09 collection. The full collection consists of roughly 1 billion web pages, comprising approximately 25TB of uncompressed data (5TB compressed) in multiple languages. The dataset was crawled from the Web during January and February 2009. Participants were encouraged to use the entire collection, however submissions over the smaller "Category B" collection of 50 million documents were accepted. Note that Category B submissions was evaluated as if they were Category A submissions. Four out of ten participating groups used the "Category B" collection.

### 3.2 Topics

Topics were defined as a typical Ad Hoc track description. To define a set of topics, we started with TREC 2009 Million Query and Web track queries and TREC 2007 Question Answering track questions. Different from the 2011 track we attempted to control two facets of search tasks as defined in [5]: "Product" and "Goal quality". The "Product" facet varied between *Intellectual* and *Factual* tasks. Intellectual tasks produce new ideas or findings (e.g. learn about a topic or make decision based on information collected), while Factual tasks only involve locating facts, data and other information items. An example of such a varietion can be viewed below.

MQ topic: 20210

Query: dehumidifiers

- Session topic: 35

  "Product" facet: Factual

  Description: You would like to buy a dehumidifier. What are some of the technical specifications you should be looking at? What is the price range for dehumidifiers? What makes one dehumidifier more expensive than another?

- Session topic: 37

  "Product" facet: Intellectual

  Description: You would like to buy a dehumidifier. On what basis should you compare different dehumidifiers?

Based on the same MQ topic, two diffent topics were developed for the Session track. The former (topic 35) describes a factual fact, with the user being asked to locate a number of facts about dehumidifiers, while the latter (topic 37) describes an intellectual task, with the user asked to

put an intellectual effort and produce a set of criteria over which different dehumidifiers could be compared.

The "Goal quality" facet varied between *specific goal(s)* and *amorphous goal(s)*. This is very similar to the dimension that [3] proposed as well-defined and ill-defined information need. Tasks with specific goals have a well-defined information need, while in tasks with amorphous goals, the information need is ill-defined. Tasks with amorphous goals might require users to redefine the topic or identify specific aspects of the subject themselves. An example can be viewed bellow:

MQ topic: 20251

Query: Swahili dishes

- Session topic: 38

  "Goal" facet: specific goals

  Description: What are some traditional Swahili dishes? What ingredients do they use to cook them? Are swahili people using any particular herb in their dishes? Could you find these ingredients in your country? Are there any recipes you can find online?

- Session topic: 40

  "Goal" facet: amorphous goals

  Description: One of your friends from Kenya invited you to attend a party in his house and have a taste of traditional swahili dishes. You would like to search and find some information about Swahili dishes.

The former topic (topic 38) defines some very specific goals while the goals of the search in the latter are not well defined. Combining "Product" and "Goal" facets for each past TREC topic we generated up to four Session topics. We can characterize a factual task with specific goals as *known-items search*, a factual task with amorphous goals as *known-subject search*, an intellectual task with specific goals as *interpretive search* and an intellectual task with amorphous goals as *exploratory search*. A complete example of all four combinations can be viewed below:

MQ topic: 20186

Query: depression symptoms

- Task: Known-item search

  Description: What is depression? What are the major symptoms of depression? What medications, therapies and other treatments can be used to treat depression symptoms? Who performs therapy and what are the costs? Does health insurance pay for any of the treatments?

- Task: Known-subject search

  Description: You think that one of your friends may have depression, and you want to search information about the depression symptoms and possible treatments.

- Task: Interpretive search

  Description: Depression is a loaded word in our culture. What are the symptoms that could differentiate depression from having just a bad month of excessive emotions? When should one seek help and what kind?

- Task: Exploratory search

  Description: A friend has been complaining for months that she is unhappy with her life. She has also mentioned that she can't easily sleep at nights. You think that she may be suffering from depression. You want to understand if this is the case and how you could assist her in getting some help from medical professionals.

Constructing topics of different task types allows to study both how user interactions different across varying task types and whether/how systems can improve search quality under different session characteristics.

### 3.3 Sessions

A session is a series of actions, including queries and clicks on ranked results, that a user performs in the process of trying to satisfy the information need represented by the topic. Through the process described above we arrived at a large set of candidate topics for the track. These topics were then presented to actual users (minus the explicit subtopics, but with the narrative including the list of questions), who would see five randomly-selected topics and asked to choose one to try to satisfy. Users then were able to use a fully-functional custom search engine for ClueWeb09 in order to satisfy the information need described by the topic.

The custom-built search interface first provided instructions to users on the tasks to be conducted (Fig. 3 in Section A). At the beginning of each session with the system, the user was shown three topics sampled randomly from the collection of topics. The user was then prompted to select one of the topics and use the search interface to satisfy the information need. The user was free to input any queries they liked, see titles and snippets for retrieved results (Fig. 4), click on URLs to see pages, and continue in this way until they determined they were finished with the topic.

The search interface used the Yahoo! BOSS (Build your Own Search Service) API to perform the actual queries, and then filtered the ranked results against the ClueWeb09 collection before they were shown to the user. This guarantees that the URLs returned are in the ClueWeb09 collection. It *cannot* guarantee that the document content is the same; in fact, it is likely that many pages have changed since ClueWeb09 was crawled. This is a compromise that we made in order to have a search system that users were likely to find satisfactory. The system requested results from Yahoo! BOSS until at least 50 documents matching ClueWeb09 URLs had been found, or 50 result pages had been requested from Yahoo! BOSS (whichever came first).

During the course of interacting with the engine, it recorded the user's interactions with the retrieval system, including the queries issued, query reformulations, items clicked, and mouse movements in the results page. Users were also asked to save web pages they clicked if they find them useful for the completion of their task. (Fig. 5). All logged information was anonymous and no identifying information about the users was saved. When the users indicated that they had fulfilled the information need, they were presented with a brief exit survey (Fig. 6) that aimed to quantify how well the search system did.

Users were mostly faculty, staff, and students at the University of Sheffield. We sent a university-wide mailing asking for participation; anyone was free to use the system. The overall approach is similar to that described by Zuccon et al. [6] for "crowdsourcing" interactions.

When data collection was complete, we had acquired a set of candidate sessions to go with the candidate topics we defined above. Each session consists of a topic, a set of queries actual users posed to Yahoo! BOSS about the topic, the returned results and the user interactions with the returned results.

We then performed some automatic and manual culling of sessions to try to achieve a set that would be interesting for the track. This involved eliminating sessions in which the user clearly didn't understand the task or the information need, eliminating sessions in which the need was satisfied after only one query, and preferring sessions with more interactions. When the culling was complete, we had a set of 98 sessions for 48 topics to release to participants.

The sessions were provided in an XML file format. An example session containing all RL4 data might look like this:

```
<session num="16" starttime="15:14:23.276482">
      <topic>
         <desc>Lara Dutta of India was crowned Miss Universe in 2000, and between 1994 and
               2000 women from India won two Miss Universe competitions, four Miss World
               competitions, and many less well-known competitions. To what extent can
               decisions and policies of the Indian government be credited with these wins?
         </desc>
         <narr>Lara Dutta of India was crowned Miss Universe in 2000, and between 1994 and
```

```
            2000 women from India won two Miss Universe competitions, four Miss World
            competitions, and many less well-known competitions. To what extent can
            decisions and policies of the Indian government be credited with these wins?
      </topic>
      <interaction num="3" starttime="15:16:33.448408">
         <query>politics 1994-2000  Indian Miss Universe</query>
         <results>
            <result rank="1">
               <url>http://en.wikipedia.org/wiki/Miss_Universe_1994</url>
               <clueweb09id>clueweb09-enwp01-55-00034</clueweb09id>
               <title>Miss Universe 1994 - Wikipedia, the free encyclopedia</title>
               <snippet>Miss Universe 1994, the 43rd Miss Universe pageant ... Later that year,
                        another Indian, Aishwarya Rai ... Malaysian Foreign Minister not to
                        make political remarks. Miss ...
               </snippet>
            </result>
            <result rank="2">
               <url>http://en.wikipedia.org/wiki/Miss_Universe_2000</url>
               <clueweb09id>clueweb09-enwp01-62-00316</clueweb09id>
               <title>Miss Universe 2000 - Wikipedia, the free encyclopedia</title>
               <snippet>Miss Universe 2000, the 49th Miss Universe pageant was held at Eleftheria
                        Stadium, ... the field they wanted to enter, be it entrepreneurship,
                        the armed forces, politics ...</snippet>
            </result>
            ...
            <result rank="10">
               <url>http://www.indianautographs.com/productdetail-216125.html</url>
               <clueweb09id>clueweb09-en0023-23-23376</clueweb09id>
               <title>Welcome to Thematic Gallery of Indian Autographs - Detailed ...</title>
               <snippet>Thematic Gallery of Indian Autographs - A ... Stylists etc. TV Stars:
                        TV Personalities: Miss India / World / Universe ... World Celebrities
                        - Political: World Celebrities ...</snippet>
            </result>
         </results>
         <clicked>
            <click num="1" starttime="15:16:43.141470" endtime="15:16:56.658945">
               <rank>2</rank>
            </click>
         </clicked>
      </interaction>
      <currentquery starttime="15:16:33.448408">
         <query>politics 1994-2000  Indian Miss Universe</query>
      </currentquery>
   </session>
```

Each experimental condition drops more data from the XML format. An RL3 session would include everything except the <clicked> blocks. An RL2 session eliminated the <results> blocks along with the <clicked> blocks. An RL1 session had virtually no information, eliminating entire <interaction> blocks.

There is a median of one reformulation prior to the last query (mean = 2.03). 19% of sessions

have three or more reformulations prior to the last query. The maximum number of queries in any session in the set is 10. There are a total of 272 recorded clicks across all sessions (2.8 per session on average). However, there are 26 sessions with no recorded clicks, and therefore an average of 3.8 clicks per session that has at least one click.

# 4 Submissions

Sites were permitted to submit up to three runs. Each submitted run includes four separate ranked result lists for all 98 sessions. Files were named "runTag.RLn", where "runTag" is a unique identifier for the site and the particular submission, and "RLn" is RL1, RL2, RL3, or RL4 depending on the experimental condition.

The track received 27 runs from the 10 groups listed in Table 1.

| | |
|---|---|
| 1. | Bauhaus-Universitt Weimar, Germany |
| 2. | Beijing University of Posts and Telecommunications, China |
| 3. | Centrum Wiskunde & Informatica (CWI), Netherlands |
| 4. | Georgetown University, USA |
| 5. | Institute of Computing Technology, Chinese Academy of Sciences, China |
| 6. | Rutgers University, USA |
| 7. | University of Albany, USA |
| 8. | University of Delaware, USA |
| 9. | University of Essex, UK |
| 10. | University of Pittsburgh, USA |

Table 1: Groups participating in the 2012 Sessions Track.

Details about the methods used by each of the participating sites can be found in the individual group's reports for the Session Track.

# 5 Session Evaluation

## 5.1 Relevance Judgments

Judging was done by assessors at NIST. As described above, each topic was the subject of one or more sessions. For each one of the 49 topics, a pool was formed from the ranked results for the past queries $q_1 ... q_{m-1}$ and the current query $q_m$ produced by Yahoo! BOSS along with the top 10 ranked documents from the submitted runs on the current query $q_m$ for all corresponding sessions.[1]

[1]For topic 27 only, because of assessing resource constraints, document were pooled at depth 5 from submissions but still at depth 10 from the Yahoo! BOSS system for the last query and previous queries.

The NIST assessors then judged each document in the pool with respect to topic description. Note that this is different from the 2011 track with document then being judged with respect to the different subtopics along with the general topic description.

The qrels produced have the following format:

```
<topic-id> 0 <doc-id> <judgment>
```

The second column used to represent the different subtopics in the 2011 track and is kept for consistency reasons. Again different from the 2011 track judgment are: -2 for spam document (i.e. the page does not appear to be useful for any reasonable purpose; it may be spam or junk.); 0 for not relevant (i.e. the content of this page does not provide useful information on the topic, but may provide useful information on other topics, including other interpretations of the same query); 1 for relevant (i.e. the content of this page provides some information on the topic, which may be minimal; the relevant information must be on that page, not just promising-looking anchor text pointing to a possibly useful page); 4 for highly relevant (i.e. the content of this page provides substantial information on the topic); 2 for key, (i.e. the page or site is dedicated to the topic; authoritative and comprehensive, worthy of being a top result in a web search engine; typically, key pages are more comprehensive, have higher quality, and are from more trustworthy sources than the merely highly relevant page); and 3 for navigational (i.e. this page represents a home page of an entity directly named by the query; the user may be searching for this specific page or site; there is often at most one page that deserves a Navigational judgment for an aspect).

The assessors were also provided with the following additional instruction: "The judgments are on the same scale of Key down to Not Relevant, and the same basic definitions of those judgment levels holds. Since an individual topic may be looking for a variety of different, though related, things, a document that covers many of those points is likely to be more relevant than a document that covers only one. But, as always, it is your opinion as the assessor that determines whether and to what extent any particular document is relevant."

Topics were assigned to assessors such that the topics that were highly related to one another (e.g., four topics on cooking Swahili foods) were all judged by the same assessor. For those topics taken from 2012 web track topics, the same assessor did both the session and web track assessing.

Relevance judgments were eventually transformed to relevance grades with spam and non-relevant documents assigned a grade of 0, relevant assigned a grade of 1, highly relevant assigned a grade of 2, key assigned a grade of 3, and navigational assigned a grade of 4.

A total of 17,861 pages were judged. Out of these 17,861 pages, 5 were judged as navigational, 458 as key, 1,360 as highly relevant, 2,679 as relevant, 12,384 as non-relevant and 975 as spam.

### 5.2 Evaluation Measures

Based on the qrels provided by NIST and the decisions described above, we evaluated the submitted runs by eight measures:

- Expected Reciprocal Rank (ERR) [2]
- ERR@10

- ERR normalized by the maximum ERR per query (nERR)
- nERR@10
- nDCG
- nDCG@10
- Average Precision (AP)
- Precision@10

## 6 Evaluation Results

Table 2 shows all results (by nDCG@10) for all submitted runs in all four experimental conditions. If RL1 (no information about the session) is the baseline, about half of the submitted runs were able to improve on that using only the information about prior queries (RL2) or using information about prior queries and retrieved results (RL3). A majority of submitted runs improved on the baseline using the interaction information (RL4). Though we cannot say for sure that RL1 is a baseline for every submission, it seems *prima facie* reasonable to conclude that the interaction information provided by RL4 can be used to improve automatic retrieval results.

Figure 1 shows changes in nDCG@10 from the RL1 baseline (left) or with increasing information (right). The three plots going down the left column show changes in nDCG@10 from using no previous data (RL1) to using greater and greater amounts of previous data. The dashed line is a difference in nDCG of zero; points above that line represent systems that saw an improvement from using the additional data while points below it represent systems that were hurt with the additional data. The 95% confidence intervals give a rough idea of whether the results are significant.

On the right-hand side, Figure 1 shows changes in nDCG@10 with increasing amounts of previous data: going from RL1 to RL2, RL2 to RL3, and RL3 to RL4. A few systems see improvement at every step. This suggests that the extra data really is beneficial for effectiveness.[2]

Tables 3 and 4 show changes in nDCG@10 when novelty in the ranking of documents for the current query is considered. In the former table documents that have been shown and clicked by the user in queries prior the current query are considered duplicates if they are also presented in the ranking for the current query and their relevance is downgraded to zero. In the latter all documents shown to the user in the queries prior to the current one are considered duplicates.

A possibility is that some systems were optimized for different aspects of effectiveness. Figure 2 shows changes in mean average precision over the RL conditions. Comparing to Figure 1 reveals some striking differences: the UvA runs tended to see large and significant improvements in MAP despite not seeing such improvements in nDCG@10, while Rutgers systems that had large and significant improvements in nDCG@10 did not see such improvements in MAP. This lends support to that hypothesis.

[2]We caution against over-interpreting this, though, as we cannot say for certain whether these conditions are comparable for every submitted run.

| run | RL1 | RL2 | | RL3 | | RL4 | |
|---|---|---|---|---|---|---|---|
| PITTSHQMsdm | **0.2615** | 0.3071 | ⇑ | 0.3103 | ⇑ | 0.3103 | ⇑ |
| **baseline** | 0.2595 | – | | – | | – | |
| PITTSHQM | 0.2558 | **0.3100** | ⇑ | **0.3221** | ⇑ | **0.3153** | ⇑ |
| PITTSHQMsnov | 0.2540 | 0.2966 | ⇑ | 0.3009 | ⇑ | 0.3019 | ⇑ |
| PITTSHQMnov | 0.2517 | 0.3009 | ⇑ | 0.3152 | ⇑ | 0.3070 | ⇑ |
| ICTNET12SER3 | 0.2481 | 0.2476 | ↓ | 0.2640 | ↑ | 0.2857 | ⇑ |
| CWIrun1 | 0.2422 | 0.2529 | ↑ | 0.2342 | ↓ | 0.2342 | ↓ |
| CWIrun3 | 0.2422 | 0.2529 | ↑ | 0.2313 | ↓ | 0.2319 | ↓ |
| gurelaxphr | 0.2334 | 0.2832 | ⇑ | 0.3033 | ⇑ | 0.2900 | ⇑ |
| guphrase1 | 0.2298 | 0.2932 | ⇑ | 0.3021 | ⇑ | 0.3021 | ⇑ |
| guphrase2 | 0.2265 | 0.2839 | ⇑ | 0.2995 | ⇑ | 0.2995 | ⇑ |
| wildcat1 | 0.2177 | 0.2130 | ↓ | 0.2715 | ⇑ | 0.2567 | ↑ |
| ICTNET12SER2 | 0.2144 | 0.2168 | ↑ | 0.2732 | ⇑ | 0.2827 | ⇑ |
| wildcat3 | 0.2068 | 0.1947 | ↓ | 0.2876 | ⇑ | 0.2608 | ⇑ |
| webis12indqe | 0.2053 | 0.2097 | ↑ | 0.2102 | ↑ | 0.2077 | ↑ |
| essexSAnchor | 0.1941 | 0.2204 | ↑ | 0.2265 | ⇑ | 0.2307 | ⇑ |
| essexSWiki | 0.1941 | 0.1899 | ↓ | 0.1899 | ↓ | 0.1899 | ↓ |
| ICTNET12SER1 | 0.1586 | 0.2043 | ⇑ | 0.2039 | ⇑ | 0.2392 | ⇑ |
| WQExpFqDSnip | 0.1515 | 0.1711 | ↑ | 0.1804 | ↑ | 0.1795 | ↑ |
| BDocExpDoc | 0.1445 | 0.1619 | ⇑ | 0.1958 | ⇑ | 0.1943 | ⇑ |
| UAlbany | 0.1407 | 0.1294 | ↓ | 0.1763 | ↑ | 0.1409 | ↑ |
| RutgersHu | 0.1272 | 0.0000 | ⇓ | 0.0000 | ⇓ | 0.1806 | ⇑ |
| RutgersM | 0.1272 | 0.0000 | ⇓ | 0.0000 | ⇓ | 0.1842 | ⇑ |
| ACombSnip | 0.1215 | 0.1240 | ↑ | 0.1801 | ⇑ | 0.1770 | ⇑ |
| webis12cnse | 0.1086 | 0.1220 | ↑ | 0.1401 | ⇑ | 0.1796 | ⇑ |
| webis12cnqe | 0.0865 | 0.1174 | ⇑ | 0.1204 | ⇑ | 0.1171 | ⇑ |
| wildcat2 | 0.0844 | 0.1338 | ⇑ | 0.2121 | ⇑ | 0.2692 | ⇑ |
| TUDrun | 0.0506 | 0.0268 | ⇓ | 0.0268 | ⇓ | 0.0268 | ⇓ |

Table 2: All results by nDCG@10 for the current query in the session for each condition (sorted in decreasing order of RL1 nDCG@10). Boldface indicates the highest nDCG@10 in the condition. $\uparrow, \downarrow$ indicate positive or negative differences from RL1. $\Uparrow, \Downarrow$ indicate statistically significant ($p < 0.05$ by a paired two-sided t-test) positive or negative differences from RL1. $\leftrightarrow$ indicates no difference from RL1. The **baseline** system is our custom search system described above.

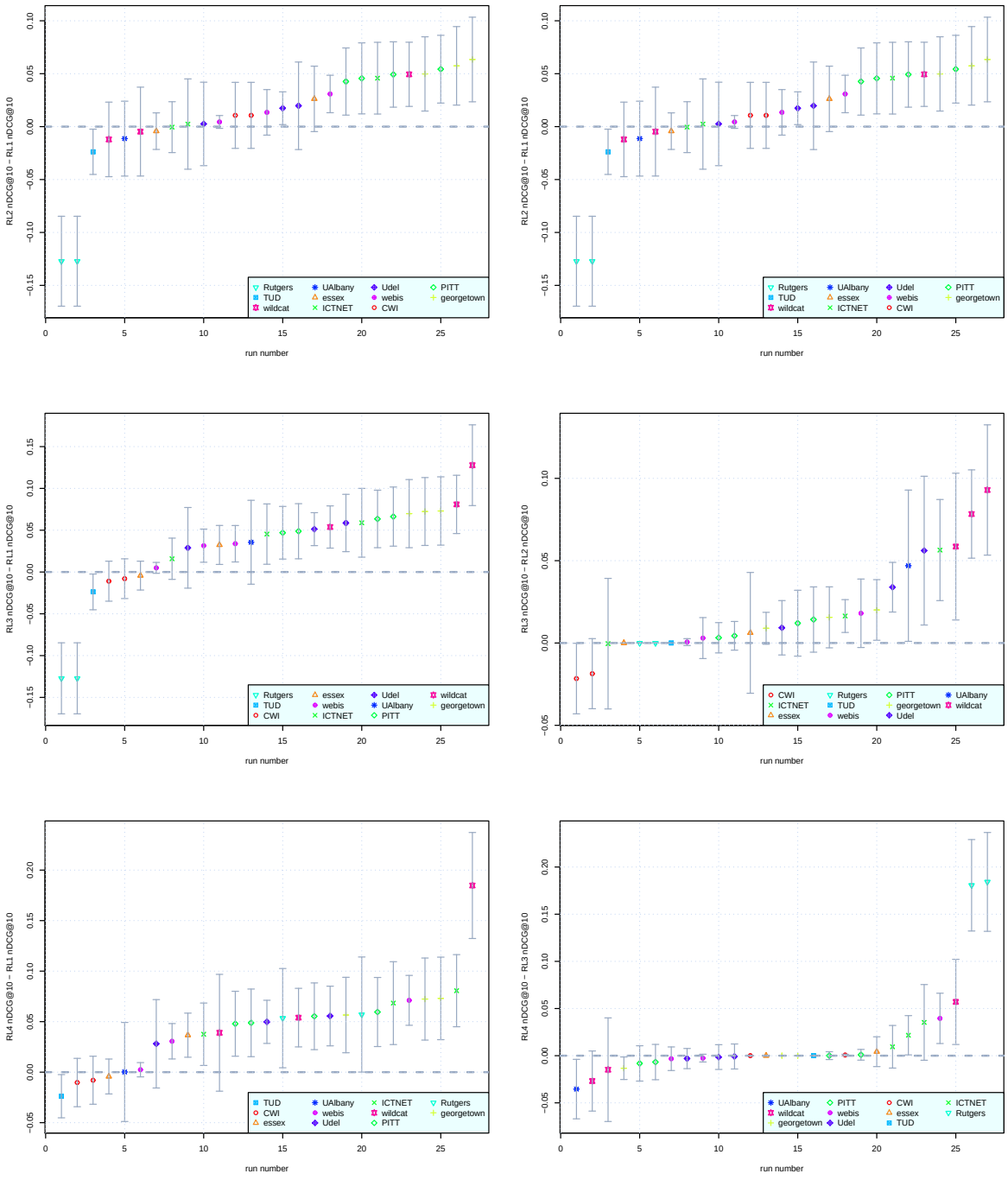


Figure 1: Left: Changes in nDCG@10 from RL1 to (from top to bottom) RL2, RL3, and RL4. Right: Changes in nDCG@10 from RL1 to RL2, RL2 to RL3, and RL3 to RL4. Error bars are 95% confidence intervals.

| run | RL1 | RL2 | | RL3 | | RL4 | |
|---|---|---|---|---|---|---|---|
| PITTSHQMsnov | **0.2540** | 0.2956 | ⇑ | 0.3009 | ⇑ | 0.3009 | ⇑ |
| PITTSHQMsdm | 0.2522 | 0.2925 | ⇑ | 0.2961 | ⇑ | 0.2953 | ⇑ |
| PITTSHQMnov | 0.2517 | **0.3009** | ⇑ | **0.3152** | ⇑ | **0.3070** | ⇑ |
| PITTSHQM | 0.2485 | 0.2955 | ⇑ | 0.3072 | ⇑ | 0.2978 | ⇑ |
| ICTNET12SER3 | 0.2401 | 0.2387 | ↓ | 0.2548 | ↑ | 0.2686 | ↑ |
| CWIrun1 | 0.2330 | 0.2426 | ↑ | 0.2257 | ↓ | 0.2257 | ↓ |
| CWIrun3 | 0.2330 | 0.2426 | ↑ | 0.2227 | ↓ | 0.2233 | ↓ |
| gurelaxphr | 0.2270 | 0.2694 | ⇑ | 0.2891 | ⇑ | 0.2768 | ⇑ |
| guphrase1 | 0.2249 | 0.2797 | ⇑ | 0.2878 | ⇑ | 0.2878 | ⇑ |
| guphrase2 | 0.2215 | 0.2701 | ⇑ | 0.2855 | ⇑ | 0.2855 | ⇑ |
| **baseline** | 0.2210 | – | | – | | – | |
| wildcat1 | 0.2086 | 0.1952 | ↓ | 0.2551 | ⇑ | 0.2253 | ↑ |
| ICTNET12SER2 | 0.2064 | 0.2087 | ↑ | 0.2639 | ⇑ | 0.2646 | ⇑ |
| wildcat3 | 0.1995 | 0.1842 | ↓ | 0.2707 | ⇑ | 0.2429 | ⇑ |
| webis12indqe | 0.1945 | 0.1989 | ↑ | 0.1994 | ↑ | 0.1969 | ↑ |
| essexSAnchor | 0.1836 | 0.2046 | ↑ | 0.2122 | ⇑ | 0.2082 | ⇑ |
| essexSWiki | 0.1836 | 0.1769 | ↓ | 0.1769 | ↓ | 0.1769 | ↓ |
| ICTNET12SER1 | 0.1556 | 0.1967 | ⇑ | 0.1973 | ⇑ | 0.2084 | ⇑ |
| WQExpFqDSnip | 0.1446 | 0.1633 | ↑ | 0.1703 | ↑ | 0.1683 | ↑ |
| BDocExpDoc | 0.1383 | 0.1548 | ⇑ | 0.1853 | ⇑ | 0.1818 | ⇑ |
| UAlbany | 0.1374 | 0.1242 | ↓ | 0.1648 | ↑ | 0.1246 | ↓ |
| RutgersHu | 0.1240 | 0.0000 | ⇓ | 0.0000 | ⇓ | 0.1507 | ↑ |
| RutgersM | 0.1240 | 0.0000 | ⇓ | 0.0000 | ⇓ | 0.1517 | ↑ |
| ACombSnip | 0.1158 | 0.1161 | ↑ | 0.1701 | ⇑ | 0.1669 | ⇑ |
| webis12cnse | 0.1038 | 0.1176 | ↑ | 0.1336 | ⇑ | 0.1744 | ⇑ |
| webis12cnqe | 0.0854 | 0.1109 | ⇑ | 0.1141 | ⇑ | 0.1106 | ⇑ |
| wildcat2 | 0.0783 | 0.1232 | ⇑ | 0.1968 | ⇑ | 0.2479 | ⇑ |
| TUDrun | 0.0492 | 0.0265 | ⇓ | 0.0265 | ⇓ | 0.0265 | ⇓ |

Table 3: All results by nDCG@10 for the current query in the session for each condition (sorted in decreasing order of RL1 nDCG@10). Clicked documents in previous queries of the session are considered duplicates and their relevance is downgraded to zero.

| run | RL1 | RL2 | | RL3 | | RL4 | |
|---|---|---|---|---|---|---|---|
| PITTSHQMnov | **0.2500** | **0.3001** | ⇑ | **0.3146** | ⇑ | **0.3063** | ⇑ |
| PITTSHQMsnov | 0.2498 | 0.2916 | ⇑ | 0.2959 | ⇑ | 0.2959 | ⇑ |
| PITTSHQMsdm | 0.2344 | 0.2650 | ⇑ | 0.2698 | ⇑ | 0.2696 | ⇑ |
| PITTSHQM | 0.2314 | 0.2746 | ⇑ | 0.2877 | ⇑ | 0.2781 | ⇑ |
| ICTNET12SER3 | 0.2192 | 0.2158 | ↓ | 0.2295 | ↑ | 0.2452 | ↑ |
| CWIrun1 | 0.2180 | 0.2240 | ↑ | 0.2088 | ↓ | 0.2088 | ↓ |
| CWIrun3 | 0.2180 | 0.2240 | ↑ | 0.2062 | ↓ | 0.2067 | ↓ |
| gurelaxphr | 0.2149 | 0.2519 | ⇑ | 0.2726 | ⇑ | 0.2611 | ⇑ |
| guphrase1 | 0.2124 | 0.2633 | ⇑ | 0.2717 | ⇑ | 0.2717 | ⇑ |
| guphrase2 | 0.2097 | 0.2526 | ⇑ | 0.2690 | ⇑ | 0.2690 | ⇑ |
| ICTNET12SER2 | 0.1936 | 0.1873 | ↓ | 0.2349 | ⇑ | 0.2419 | ⇑ |
| wildcat1 | 0.1843 | 0.1682 | ↓ | 0.2269 | ⇑ | 0.2082 | ↑ |
| wildcat3 | 0.1811 | 0.1633 | ↓ | 0.2441 | ⇑ | 0.2188 | ⇑ |
| webis12indqe | 0.1665 | 0.1714 | ↑ | 0.1720 | ↑ | 0.1681 | ↑ |
| essexSAnchor | 0.1662 | 0.1822 | ↑ | 0.1892 | ⇑ | 0.1905 | ⇑ |
| essexSWiki | 0.1662 | 0.1608 | ↓ | 0.1608 | ↓ | 0.1608 | ↓ |
| ICTNET12SER1 | 0.1388 | 0.1724 | ⇑ | 0.1781 | ⇑ | 0.1899 | ⇑ |
| WQExpFqDSnip | 0.1309 | 0.1483 | ↑ | 0.1536 | ↑ | 0.1527 | ↑ |
| **baseline** | 0.1300 | – | | – | | – | |
| BDocExpDoc | 0.1242 | 0.1407 | ⇑ | 0.1670 | ⇑ | 0.1646 | ⇑ |
| UAlbany | 0.1241 | 0.1098 | ↓ | 0.1436 | ↑ | 0.1142 | ↓ |
| RutgersHu | 0.1079 | 0.0000 | ⇓ | 0.0000 | ⇓ | 0.1344 | ↑ |
| RutgersM | 0.1079 | 0.0000 | ⇓ | 0.0000 | ⇓ | 0.1411 | ↑ |
| ACombSnip | 0.1054 | 0.1064 | ↑ | 0.1534 | ⇑ | 0.1512 | ⇑ |
| webis12cnse | 0.0885 | 0.1087 | ⇑ | 0.1145 | ⇑ | 0.1214 | ⇑ |
| webis12cnqe | 0.0804 | 0.0955 | ⇑ | 0.0952 | ↑ | 0.0939 | ⇑ |
| wildcat2 | 0.0707 | 0.1064 | ⇑ | 0.1708 | ⇑ | 0.2248 | ⇑ |
| TUDrun | 0.0431 | 0.0248 | ⇓ | 0.0248 | ⇓ | 0.0248 | ⇓ |

Table 4: All results by nDCG@10 for the current query in the session for each condition (sorted in decreasing order of RL1 nDCG@10). Shown documents in previous queries of the session are considered duplicates and their relevance is downgraded to zero.

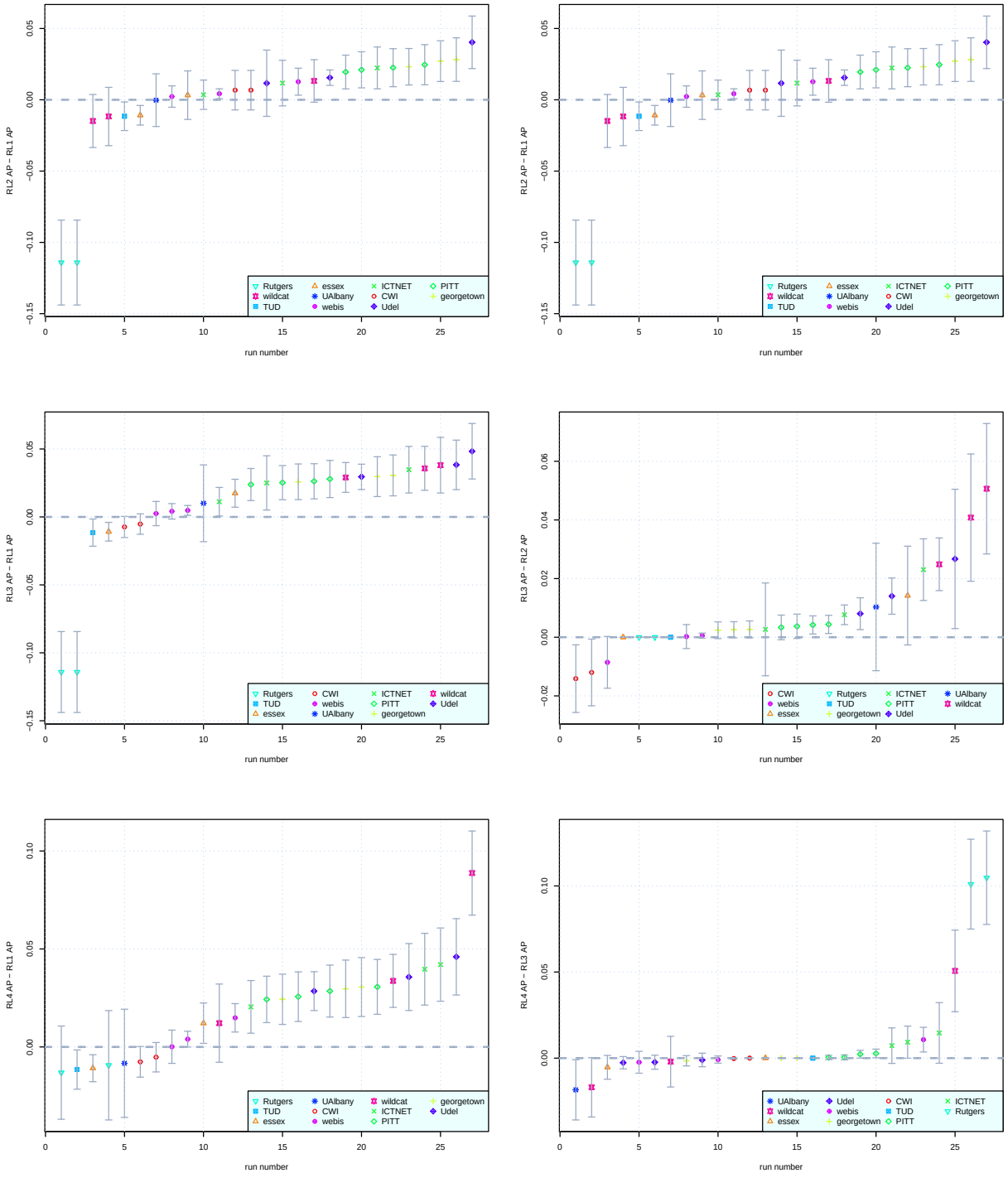


Figure 2: Left: Changes in MAP from RL1 to (from top to bottom) RL2, RL3, and RL4. Right: Changes in MAP from RL1 to RL2, RL2 to RL3, and RL3 to RL4. Error bars are 95% confidence intervals.

## References


[1] B. Carterette, E. Kanoulas, P. D. Clough, and M. Sanderson, editors. *Proceedings of the ECIR 2011 Workshop on Information Retrieval Over Query Sessions*, Available at `http://ir.cis.udel.edu/ECIR11Sessions`.

[2] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *In Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM)*, 2009.

[3] P. Ingwersen and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[4] E. Kanoulas, B. Carterette, M. Hall, P. Clough, and M. Sanderson. Session track 2011 overview. In *The Twentieth Text REtrieval Conference Proceedings (TREC 2011)*. National Institute of Standards and Technology, 2012. (`http://trec.nist.gov/pubs/trec20/papers/SESSION.OVERVIEW.2011.pdf`).

[5] Y. Li and N. J. Belkin. A faceted approach to conceptualizing tasks in information seeking. *Inf. Process. Manage.*, 44(6):1822–1837, Nov. 2008.

[6] G. Zuccon, T. Leelanupab, S. Whiting, E. Yilmaz, J. Jose, and L. Azzopardi. Crowdsourcing interactions: Capturing query sessions through crowdsourcing. In Carterette et al. [1].

# A Screenshots of the Search Interface

**ClueWeb Search Instructions**
A set of tasks will appear on the next screen. Please read the tasks carefully and select the task that you are most familiar with and understand best.

Please do not select any tasks that you have already completed. If you have completed all the tasks that are listed, then please refresh the page for a new list of tasks.

After clicking on the "Select this task" button for the task that you wish to select, a search box will appear. Type in the keywords that you believe are the best to find web pages that will help you fulfil the selected task. Please refer to the task description on the left-hand side any time you need to review the task.

Use the search system naturally, as you would do in your everyday searching activities. Browse the ranked list of web-pages, click on any web-page that you think may be useful, click on the next page button if you want to see more results, or reformulate your search query and search again if you think that the returned results are not satisfying.

Any web-page you click on will be opened in a new browser window / tab. Please only view the content on that web-page and do not click on any links on that web-page. After you have viewed the web-page, please close the new window / tab and return to the ClueWeb search results. You must then select whether you would save that web-page in order to solve your task or not.

You can finish the task by clicking on the "Finish" button in the left-hand side-panel. You may finish at any time if you are satisfied with the web pages you have observed so far.

After you finish a task, you will be shown a short questionnaire. Please fill this out and click on the "Save" button. You will then be taken back to the task selection page and can select another task to do. Please do as many tasks as you can in the time that you have.

Note that your browsing activities will be recorded. The recored data is anonymised and then stored securely. It will be used for research purposes only.

Note that the search engine is still experimental and thus slower than most current search engines and will also return fewer results.

There is a help icon (?) in the top-right corner of every page that will show these notes at any time.

If you have any further questions regarding the tasks, please contact ekanoulas at gmail dot com. If you find a bug or the system just doesn't work for you, please contact m dot mhall at sheffield dot ac dot uk.

Figure 3: The instruction text used for the data-generation task

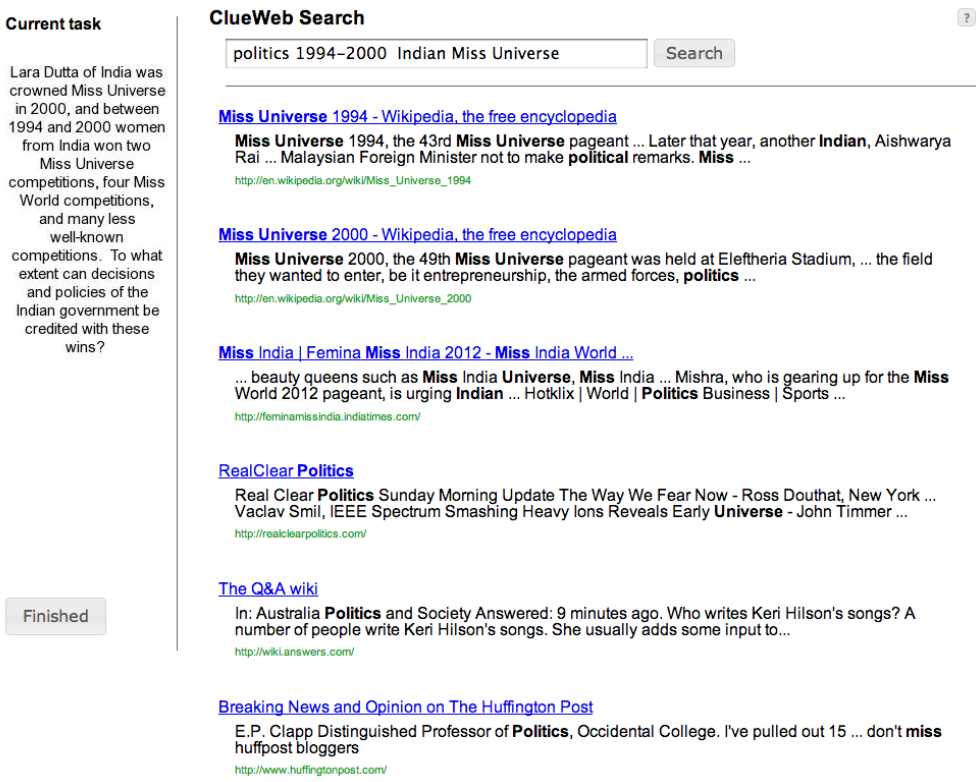


Figure 4: The result presentation UI.

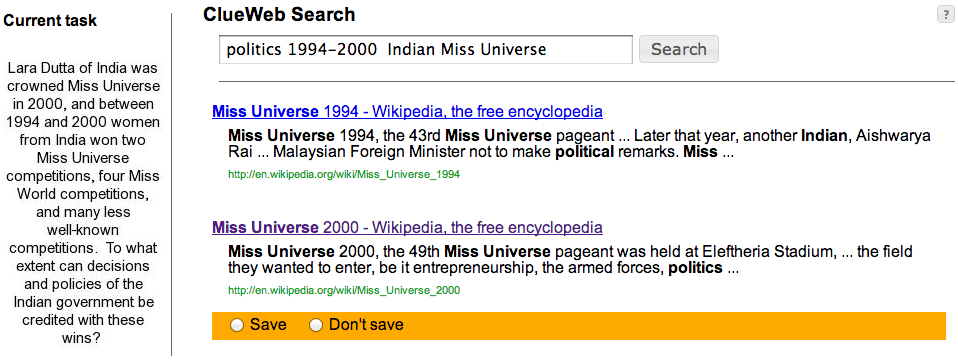


Figure 5: The page rating UI.

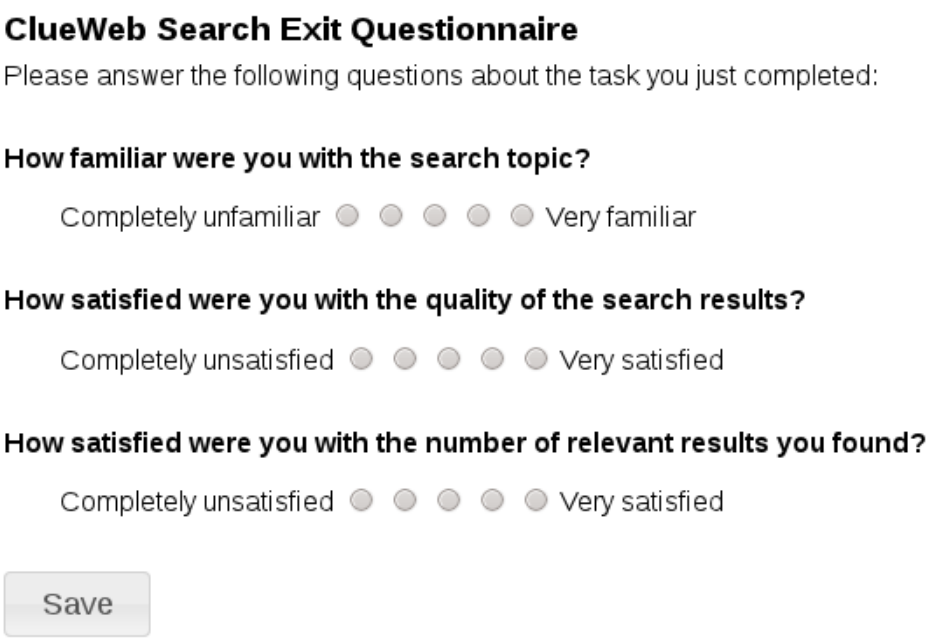


Figure 6: The exit interview UI