AutoAdapt @ TREC 2010

Dyaa Albakour

October 7, 2010

Table of contents

1. The AutoAdapt Project
2. TREC 2010
   - What is TREC?
   - The Session Track
3. ClueWeb09 and Indexing
4. Experiments
   - Overview
   - Baseline 1
   - Baseline 2
   - The AutoAdapt Approach
5. Future Work

Update on the AutoAdapt Project

- Ant Colony Optimisation for Deriving Suggestions from Intranet Query Logs, WI10 paper.
- A Methodology for Simulated Experiments in Interactive Search. SimInt 2010 @ SIGIR.
- Towards Adaptive Search in Digital Libraries. Submitted as a book chapter for AT4DL.
- Building an adaptive search system. Collaborating with a number of Industrial partners.

What is TREC?

- The purpose was to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies.

What is TREC?

- The purpose was to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies.
- Co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense. started in 1992.

What is TREC?

- The purpose was to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies.
- Co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense. started in 1992.
- Annual Competition:
  Tracks announced in February.
  Results usually submitted in summer.
  Assessments are back in September.
  Conference takes place November.

What is TREC?

- The purpose was to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies.
- Co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense. started in 1992.
- Annual Competition:
  Tracks announced in February.
  Results usually submitted in summer.
  Assessments are back in September.
  Conference takes place November.
- seven tracks in TREC 2010: Blog Track, Chemical IR track, Entity Track, Legal Track, Relevance Feedback track, Session track, Web Track.

The Session Track

- Evaluate the effectiveness of search engines in interpreting query reformulations.

The Session Track

- Evaluate the effectiveness of search engines in interpreting query reformulations.
- A good search engine should be able to utilise the previous queries in the sequence of a session to provide better results that reflect the user needs throughout the session.

The Session Track

- Evaluate the effectiveness of search engines in interpreting query reformulations.
- A good search engine should be able to utilise the previous queries in the sequence of a session to provide better results that reflect the user needs throughout the session.
- Example:
  Britney Spears → Paris Hilton
  France Hotels → Paris Hilton

The Session Track

- Evaluate the effectiveness of search engines in interpreting query reformulations.
- A good search engine should be able to utilise the previous queries in the sequence of a session to provide better results that reflect the user needs throughout the session.
- Example:
  Britney Spears → Paris Hilton
  France Hotels → Paris Hilton
- The session track provides a framework to assess this particular issue in Information Retrieval systems.

The Session Track

- Evaluate the effectiveness of search engines in interpreting query reformulations.
- A good search engine should be able to utilise the previous queries in the sequence of a session to provide better results that reflect the user needs throughout the session.
- Example:
  Britney Spears → Paris Hilton
  France Hotels → Paris Hilton
- The session track provides a framework to assess this particular issue in Information Retrieval systems.

The Session Track - The Task

- Only sessions with two queries are considered this year.
- Participants are given a set of 150 query pairs, each query pair (original query, query reformulation) represents a user session.
- The participants are asked to submit three ranked lists of documents form the **ClueWeb09** dataset:
  - One for the original query ($RL1$).
  - One for the query reformulation ignoring the original query ($RL2$).
  - One for the query reformulation taking the original query into consideration ($RL3$).

The Session Track - Type of Queries

1. **Generalisation**: ‘low carb high fat diet’ → ‘types of diets’.
2. **Specification**: ‘us map’ → ‘us map states and capitals’
3. **Drifting/Parallel Reformulation**: ‘music man performances’ → ‘music man script’.

The Session Track - Evaluation

1. Can search engines improve their performance for a given query using previous queries? $RL2, RL3$
2. How do they perform over an entire session? $RL1, RL3$.

- $PC(10)$ and $nDCG(10)$ will be exactly estimated.
- Participants can be ranked and their performance can be compared over $RL2$ and $RL3$.
- Primary comparison measure between participants is the $nDCG(10)$ for $RL3$.
- Documents that appear in $RL1$ will be penalised if they reappear in $RL2$ and $RL3$.

The ClueWeb09 Dataset

- 1,040,809,705(1 billion) web pages, in 10 languages.
- ClueWeb09 Category B: 50m English pages (Tier 1 web crawl).
- Public index available using Indri Search Engine .
- The Indri search engine supports language retrieval models (query likelihood model).

The Runs Matrix

	RL1	RL2	RL3
System 1	D_q	Dr	(baseline 1)
System 2	D_q	Dr	(baseline 2)
System 3	D_q	Dr	(AutoAdapt Approach)

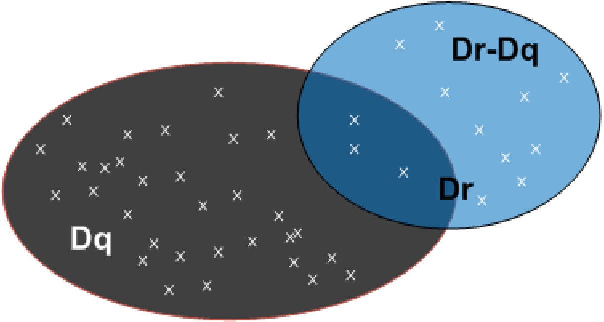
- q: The original query consisting of a number of terms qt_i.
- r: The reformulated query consisting of a number of terms rt_i.
- Dq: a ranked list of documents returned by Indri
  $D_q < d_{q,1}, d_{q,2}, ..., d_{q,n} >; d_{q,i} \notin SPAM, n < 1000$
  Query likelihood model.
- 70% of ClueWeb09 documents are considered spam.

Baseline 1

- For $(RL3)$, we return the list D_{q+r}:
  Submit a query $qt \cup qr$.
  Indri combine function
  becoming dj $\rightarrow$ dj jobs
  Submitted Indri query: combine(becoming dj jobs)

Baseline 2

- For $(RL3)$, we return the list:
  $D_r - D_q = \{d; d \in D_r, d \notin D_q\}$
  The documents in $D_r - D_q$ are ordered using their ranking in D_r



Mining Query Logs

- Fonseca's Association Rules from query logs to extract query suggestions [3].

Mining Query Logs

- Fonseca's Association Rules from query logs to extract query suggestions [3].
- Ant Colony Optimisation [2] to learn query suggestions from Intranet query logs.
- Can we approximate the user session to the graph extracted from query logs?

Mining Query Logs

- Fonseca's Association Rules from query logs to extract query suggestions [3].
- Ant Colony Optimisation [2] to learn query suggestions from Intranet query logs.
- Can we approximate the user session to the graph extracted from query logs?
- Possible Solution:
  1. Extract associations for both queries in the session.
  2. Expand the reformulated query with the intersection of suggestions extracted for both queries.

Anchor Logs to mimic Query Logs

- Anchor log as a simulated query log has been shown to be effective in query reformulation Dang and Croft, WSDM10[1].
- The anchor log from ClueWeb09, the University of Twente[4].

  - Anchor log from ClueWeb09 Cat B, 3 GB.
  - Anchor Text for about 87 % of the documents, 43m lines
  - (TREC-ID, URL, ANCHOR TEXT)
  - Example:

```
clueweb09-en0000-23-00060
http://001yourtranslationservice.com/dtp/
‘website design’ ‘DTP and Web Design’
‘Samples’ ‘programmers’ ‘desktop publishing’ ‘DTP pages’ ‘DTP samples’
‘DTP and Web Design Samples’ ‘DTP and Web Design Samples’ ‘DTP and Web Design
Samples’ ‘DTP and Webpage Samples’ ‘DTP’

http://001yourtranslationservice.com/dtp/
```

Experimental steps

- Remove all the stop words from both queries in the session.
- Extract all the lines (the sessions) in which the anchor text contains either queries.
- Fonseca's Association rules [3] to extract all the suggestions for both constituents of the session pair.
- Consider the top 10 phrases or terms in the set composed by the intersection of the suggestions extracted for both constituents as useful expansions to the reformulated query plus the original query.

Examples

Session	**Expansion terms or phrases**
gps devices → 'garmin'	'gps devices', 'wikipedia','usb', 'gps device', 'gps products', 'garmin nuvi880', 'garmin gps device','visit garmin'
computer worms → malware	'computer worms','computer security', 'category','worm'
us geographic map → us political map	'us political map','article'

Future Work

- Classifying the sessions.
- Indexing the entire ClueWeb09 collection in house.
- Analysis of the results when assessments are received.
- The availability of relevance judgements would help us to improve our method and try out new approaches in the lab.

V. Dang and B. W. Croft.
Query reformulation using anchor text.
In *WSDM '10: Proceedings of the third ACM international conference on Web search and data mining*, pages 41–50, New York, NY, USA, 2010. ACM.

S. Dignum, U. Kruschwitz, M. Fasli, Y. Kim, D. Song, U. Cervino, and A. De Roeck.
Incorporating Seasonality into Search Suggestions Derived from Intranet Query Logs.
In *Proceedings of the IEEE/WIC/ACM International Conferences on Web Intelligence (WI'10)*, pages 425–430, Toronto, 2010.

B. M. Fonseca, P. B. Golgher, E. S. de Moura, and N. Ziviani.
Using association rules to discover search engines related queries.

In *Proceedings of the First Latin American Web Congress*, pages 66–71, 2003.

D. Hiemstra and C. Hauff.
Mirex: Mapreduce information retrieval experiments.
Technical Report TR-CTIT-10-15, Centre for Telematics and Information Technology University of Twente, Enschede, April 2010.