

Query Log Analysis Workshop 2009

Paul Clough

Department of Information Studies
University of Sheffield

Query logs

- Common for many online systems to record interactions between people using the system and responses from the system
 - Visitor's queries and clicks and system responses is known as *click-stream data*
- Offers potentially valuable information for a wide range of applications
 - E.g. design, personalisation, evaluation of websites and search systems
- Many organisations collect log data
 - But what data do they collect?
 - How are they using the data?
 - What do they want to do with the data?

Query log analysis

- Objectives of a log analysis (Mat-Hassan & Levene, 2005)
 - To investigate a searcher's performance
 - To establish the profile of an effective searcher
 - To establish a user's searching characteristics
 - To understand a user's navigational behaviour (including the number of query terms entered and the number of click-throughs viewed)
- Provides a new paradigm for evaluating search
 - Major search engine companies exploit logs

Query log analysis

- Lots of research in different communities
 - Information seeking and retrieval
 - Web usage mining
 - Web analytics
 - Business and marketing
 - Personalisation and adaptive systems
- But much of the work is disjoint and not shared between communities
 - Limited synthesis of research (e.g. terminology and approaches)
 - Lack of standardised procedures and resources

Query log analysis workshop

- Aims of the workshop
 - To provide a forum in which researchers can discuss the current activities in log analysis and stimulate ideas on the future directions and challenges facing the field
- An event to attract researchers and practitioners
 - Providing an environment in which to brain-storm and discuss
- Enabled a bringing together of people doing (or interested in) log analysis
 - Invited speakers from both industry and academia
 - Further attendees interested in log analysis
- In summary
 - What research has been done, where is it now, where is it going?



Treble-CLEF



The CLEF research results have led to development of a new generation of multilingual retrieval system prototypes

BUT lack of technology transfer

Treble-CLEF extends the CLEF activity by:

- continuing to promote MLIA R&D via evaluation campaigns;
- providing a consistent training activity: tutorials, workshops, summer school;
- producing best practice guidelines for system implementation;
- providing resources to encourage the multilingual system development.

www.trebleclef.eu



Approach



- ♦ **Evaluation**
 - test collections and laboratory evaluation
 - user evaluation
 - log analysis
- ♦ **Best Practices & Guidelines**
 - system-oriented aspects of MLIA applications
 - collaborative user studies
 - user-oriented aspects of MLIA interfaces
- ♦ **Dissemination and Training**
 - tutorials
 - workshops
 - summer school

Task 4.1 – what the proposal says

- Task 4.1 of WP4 (Evaluation Methodologies)
 - Organisation of a **brain-storming** workshop to bring together **researchers** to share their **experiences** in using log analysis with the aim of defining best practices
 - **Results** will be actively disseminated to the DL community

Format

- Planning to hold an event to attract researchers and (possibly) practitioners
 - Mainly an environment in which to brain-storm and discuss
- Invite some well-known academics from different communities to speak on log analysis
 - Set some questions for people to discuss
 - Maybe submit short position paper (for publication)
 - Cover their expenses
- Investigating whether we can hold an event in collaboration with the British Computer Society (BCS)
 - Would provide route for marketing the event (attract further interested academics and practitioners)
 - BCS has conference facilities in London
 - Initial contact with Andy MacFarlane (City University)

Wednesday 27th May 2009 (Wilks room 2, BCS London Office)

9.30-10.00	Arrival (tea and coffee)	
10.00-10.30	Welcome and introduction	Paul Clough (University of Sheffield)
10.30-11.00	Presentations	Mark Levene (Birbeck, University of London)
11.00-11.30		Nigel Ford (University of Sheffield)
11.30-12.00		Jim Jansen Penn State University)
12.00-13.00	Lunch	
13.00-13.30	Presentations	Filip Radlinski (Microsoft Research)
13.30-14.00		Vanessa Murdock (Yahoo! Research)
14.00-14.30		Lynn Connaway (OCLC)
14.30-15.00		Dhaval Thakker (Press Association)
15.00-15.30	Break	
15.30-16.00	Presentations	Fabrizio Silvestri (ISTI-CNR)
16.00-16.30		Bettina Berendt (Katholieke Universiteit, Leuven)
16.30-17.00		Udo Kruschwitz (University of Essex)
20.00	Workshop Dinner	Zizzi Ristorante (73-75 The Strand, WC2R 0DE) http://www.zizzi.co.uk/restaurants/93

Thursday 28th May 2009 (Wilks room 2, BCS London Office)

08.30-9:00	Arrival (tea and coffee)	
09.00-10.00	Tutorial session	Jim Jansen (Penn State University)
10.00-10.30	Presentations	Giorgio Di Nunzio (University of Padoa)
10.30-11.00		Thomas Mandl (University of Hildesheim)
11.00-12.00	Discussion	(see discussion questions)
12.00-13.00	Lunch	
13.00-13.30	Discussion	(see discussion questions)
13.30-14.00		
14.00-14.30		
14.30-15.00		
15.00-15.30	Summary and close	Paul Clough (University of Sheffield)
15.30	Depart	

Discussions

- Event is planned to last one day (longer, 1.5, 2?)
 - Date [TBC]
- Morning session
 - Short talks from participants
 - Share experiences and stimulate points for discussion
- Lunch
- Afternoon session
 - Discussion to address workshop questions (structured brain-storming, could be longer?)
 - Maybe divided into groups
- Extra
 - Hands on work?

Discussion questions

- What approaches to log analysis are used in different fields?
- What are problems with log analysis in different fields?
- Which techniques are similar between fields/applications? (Which techniques are specific to particular applications?)
- How can we effectively transfer research into industry?
- How can researchers get access to logs? (E.g. what will stop industry from sharing logs?)
- What approaches could be used to generate logs to share within the research community?
- How generalisable are the techniques/findings of log analysis on specific logs?

Discussion questions

- How can we evaluate approaches to log analysis? (What kind of benchmarks do we need, how do we generate them and what kind of evaluation campaign should be run?)
- What are the future challenges/directions for the field of query log analysis/mining? (e.g. eye tracking, web search advertising, time-series analysis of queries, integration of multiple transaction logs, correlating transaction logs with user behaviour)
- How can we bring researchers from different disciplines closer together?
- What are the niches and contributions that academia can make to log analysis?
- Where are areas for academic - industry collaboration?
- How can we generate funding opportunities from grant agencies in log analysis?
- Can we develop a meta-methodology that combines log analysis with other methods to provide a "truer" picture of the user - system - information interaction process?

Questions to ask SMEs

- Do you collect query logs?
 - How, frequency, which system, for what purposes?
- Do you mine/analyse query logs?
 - What Web Analytics tools do you use?
 - What do you do with them?
 - Do they fulfil your needs?
 - What problems do you face?
- Do you gather logs from your search systems?
- What potential do you see from mining your logs?
- What problems do you face with analysing your logs?
- Do you perform query or click-through analysis?