Abstract
This article briefly characterises the fields of Question Answering and Web Log Analysis, and summarises the main achievements and research methods of each. It then goes on to discuss links between these fields and to describe possible research questions which could be investigated. Finally, it summarises the content of the accepted papers at the workshop and relates these papers to the research questions.

1. Introduction
A Question Answering (QA) system takes as input a short natural language query and gives back an exact answer to it, extracted from a document collection (Hirschman and Gaizauskas, 2001). The origins of QA under this definition can be traced back to TREC-8 (Voorhees and Harman, 1999) when the first QA task was organised. Since then, there have been numerous conferences and workshops concerned with this important field.

A Web Log is a record of a person’s internet search; invariably it specifies what query they typed in, and it may include many other pieces of related information such as what sites they subsequently visited, how they subsequently modified their query and so on. Web Log Analysis (WLA) is the field which seeks to examine logs by various means and to use the information gained to improve search engines. One of the earliest studies in this area was Jansen, Spink and Saracevic (2000).

What, then, have these fields in common and what can they learn from each other? Will the fields converge, and what will be the next issues to investigate? The purpose of the Web Logs and Question Answering (WLQA) workshop is to answer these and related questions. In this introductory article we adopt the following strategy. Firstly we give a brief overview of QA and WLA, outlining their history and the key techniques used. Secondly, we state briefly what we consider to be the key areas which are common to both, the principal research questions in these areas, and the means by which they might be investigated. In the final section we summarise the contents of the workshop papers and attempt to fit them into the above picture.

2. Question Answering
2.1 Aims and Background
As understood in the modern sense, a QA system takes as input a short natural language question and produces as output an exact answer, extracted from a document collection. The first time systems of this kind were developed or discussed was probably at the Text REtrieval Conference (TREC) where a QA track was introduced at TREC-8 (Voorhees and Harman, 1999). Up until that time, TREC was mostly concerned with the development and evaluation of Information Retrieval (IR) systems using a common framework for the evaluation and a shared document collection. The QA track inherited these ideas, the key difference being that in QA an exact answer was expected while in IR a list of documents was produced. In order to achieve this exactness, a vital idea was that of the Named Entity (NE) - a piece of information such as a person name, company name, time, date, place etc which could be recognised in a text and hence become a candidate exact answer. The NE idea had been recently developed and used at the Message Understanding Conference evaluations in Information Extraction beginning in 1987 (Grishman, 1996).

In passing we should mention that QA under other definitions goes back to the earliest days of artificial intelligence research. For example, Simmons (1965) discusses fifteen different systems. A full history can be found in Hirschman and Gaizauskas (2001).

Since TREC, QA research has become a very active area and as a result of this work, highly sophisticated systems such as TrueKnowledge (2010) and WolframAlpha (2010) have started to appear.

2.2 Key Techniques in QA
While different approaches to QA have been taken, many systems have converged on a standard architecture. Prager (2006) gives three components: Question Analysis,
Search and Answer Extraction. Other authors give more components by subdividing these further, but the essence is the same.

During Question Analysis, the type of question and the form of NE expected as the answer are determined. For example if we have ‘Who is the president of the USA?’ then the type of the question is WHO and the expected answer type (i.e. NE) is PERSON. During the Search stage, the document collection is scanned for documents which might contain the answer to the question. The simplest approach to this task involves creating and submitting an IR query based on keywords derived from the question. Finally, in Answer Extraction, candidate NEs of type PERSON are selected from likely documents, and the best one is returned as the answer.

Concisely stated, a number of techniques have proved valuable and replicable in pursuit of the above stages. Firstly, NE recognition is a key process because it must be highly accurate and avoid confusions, for example between person names and company names which are often identical. Initially, approaches revolved around the use of lists (e.g. country names etc) or grammatical analysis (e.g. for dates and times which have an internal structure). In recent years, however, attention has shifted to machine learning approaches which generalise their knowledge from a set of training examples.

Secondly, Predictive Annotation (Prager et al., 2000) allows documents to be retrieved in the Search phase which are guaranteed among other things to contain an NE of the required type. This is done by including pseudo-words in the source documents such as SNAME which are then indexed by the search engine and can be used at query time.

Early Answering (Clarke, 2002) was a reaction to certain TREC questions which asked about Baseball scores, State Flowers and other standard pieces of information which could be determined in advance without looking at the documents at all. The solution was to organise such information in tables and then find the answer directly.

Another key idea, was the logical connection within a text of key elements, rather than their simple co-occurrence (Moldovan et al., 2008). Inference could be based on knowledge structures e.g. derived from WordNet (Fellbaum, 1998) or achieved by word chaining. Related to this was Answer Validation - checking an answer following extraction to see which it was likely to be correct. An influential approach was that of Magnini et al. (2002) which exploited the redundancy of the web.

Another approach to the use of redundancy in QA was to combine multiple ‘pipeline’ architectures of the type outlined above into one system (Ahn et al., 2005). This allowed several different mechanisms to be used for producing answer candidates, and several to to be used for scoring them. For an extensive review of QA, see Prager (2006).

Finally, it should be mentioned that a ‘second generation’ approach to QA involves the representation of documents in a structured and analysed form, achieved during indexing rather than at query processing time. Examples of such systems are START (2010) from MIT, and Powerset (2010).

2.3 Strengths and Weaknesses of QA

In the above we have attempted to summarise activity in a large field over a number of years. What then are the key achievements and shortcomings of all this work? First of all, the tracks at TREC and their closely related counterparts at CLEF (Peñas et al. 2010) and NTCIR (Mitamura et al. 2008) have resulted in a wide understanding of how to build efficient and effective QA systems of the type needed to answer simple factoid questions and others of closely related types. Moreover, work at CLEF and NTCIR has shown that these ideas can be transferred over very effectively to monolingual QA systems in languages other than English.

However, there are also significant weaknesses. The range of questions has been extremely narrow and these have been asked against fixed document collections because this is convenient and practical rather than realistic. In addition, the evaluation paradigm incorporating the judgements Right-Wrong-ineXact-Unsupported has been widely followed for the same reasons. In consequence, QA has in the main tended to ignore the questions of real users or the importance of dialogue as a fundamental basis for answering real questions. Instead, questions have been back-enabled from convenient sources such as newspapers or the Wikipedia.

Due to significant performance improvements at TREC over the years, QA has come to be regarded as a solved problem where no more research can be usefully conducted. However, monolingual factoid QA is only a small part of the overall question answering problem whose solution is essential to the aim of making machines more usable. (Cross-lingual QA has also shown to be a completely unsolved problem.) What is needed therefore are new ideas and new directions. This is the key rationale for the current workshop.

3. Query Log Analysis

3.1 Aims and Background

A query log is a record of queries entered into an internet search engine, together in some cases with additional information such as sites visited etc. According to Clough (2009) quoting Mat-Hassen and Levene (2005), the objectives of Query Log Analysis (QLA) are

- 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.

A key starting point for QLA was a panel entitled ‘Real life information retrieval: a study of user queries on the Web’ at SIGIR in 1997. A landmark paper appeared as a follow-up: Jansen, Spink and Saracevic (2000). The focus of this paper was a log of the Excite engine containing 51,473 entries. The authors conducted a manual analysis of sessions, queries, terms within queries, characteristics of the user and an analysis of failed queries to identify trends within user mistakes. Since that paper there has been a growing interest in QLA.

LogCLEF was first evaluation campaign track focusing on logs. It started at CLEF 2009 (Mandl et al., 2010). The goals were to understand search behaviour especially in multilingual contexts and, ultimately, to improve search
systems. There were two tasks, geographic query identification and library search use. For the former task, logs were obtained from the Tumba! search engine and from the European Library search engine. The purpose was to identify geographical entities in the queries. For the latter task, just the library log was used. Each group carried out a different investigation, including finding translations of queries, searching for documents via search terms previously used to find them, query re-formulation, and analysing characteristics of sequences of queries.

Prior to LogCLEF there was also a log-file based task within GeoCLEF 2007 which used the MSN query log (Mandl et al. 2008).

In 2009, Jansen, Spink and Taksa published a comprehensive handbook on QLA, summarising much of the research which has so far been conducted.

Also in 2009, an important workshop took place entitled Query Log Analysis (Clough, 2009). Some of the main techniques in current use were described there.

Finally, new conferences devoted to QLA and related topics have been established, including WSDM (2008) and WSCD (2009).

3.2 Key Techniques of QLA

Some key approaches to QLA are outlined here. Firstly, the most fundamental form of analysis gathers information such as the numbers of sessions, terms etc and the production of statistics based on these. Jansen et al. (2000) was mainly of this type, and many papers have followed.

Secondly, there have been manual analyses of queries in small numbers looking for detailed aspects such as query type or focus. These sorts of studies have been highly informative but they are limited in the number of queries which can be examined.

Thirdly, there have been automatic analyses which nevertheless do not use Machine Learning (ML) algorithms. For example, Bernstram, Herskovic and Hersch (2009) categorise queries by mapping terms within them onto a domain specific ontology.

Fourthly, ML algorithms have been adopted to carry out tasks such as query classification. Here are several examples of this type of work.

Taksa, Zelikovitz and Spink (2009) show how short queries can be categorised by exploiting information gleaned using IR techniques, a method previously used by Sarawagi (2005).

For topic identification, Ozmutlu, Ozmutlu and Spink (2009) identify three classes of supervised ML algorithm which are effective: Maximum Entropy models, Hidden Markov models and Conditional Random Fields.

Levene (2009) advocates the use of Support Vector Machines (supervised ML) for the classification of queries. He also points out that queries need enriching with result pages or snippets, with related queries and with a training set of categorised web pages. Another line of work has been the prediction of the next link which a user will follow, using Markov chains constructed from logs.

Fifthly, there has been a focus on clickthrough data in the context of web search. Some logs specify what URLs a user clicked when they were returned in response to a query by a search engine. These tell us something important, most obviously that the user was interested enough in a link to click on it. Studies focus on the relation of query and page content, the time spent on result pages and the behaviour on the search engine result page. Radlinski et al. (2008) is one of many studies concerned with the analysis of such data. Murdock et al. (2009) also use clickthrough data to try to predict which is the best advertisement to show following a query in Yahoo!. They use a Binary Perceptron for this task.

Sixthly, in parallel with the above, there have been a number of studies involving people who participate in a carefully designed experimental study in query formulation etc. In all such cases the logs are captured though the numbers of queries involved tends to be small. For an extensive review of QLA, see Silvestri (2010).

3.3 Strengths and Weaknesses of QLA

The key strength of QLA is that there is potentially a huge amount of data available which is being generated in a naturalistic way, without the users being aware that they are being monitored. This differs greatly from most QA work so far, where queries are generated manually and are therefore not naturally occurring.

The main weakness of QLA is perhaps this same point - the huge amount of data. It cannot be analysed manually and because of its relatively sparse nature (just the query typed, the sites visited etc) we can never know for certain what a user really intended. This can only be inferred, and not with certainty. We do observe user behaviour in log analysis but at much lesser detail than we could observe it in a test environment. Information need, satisfaction and opinion about result documents can only be guessed from logs.

Another difficulty is that search engine companies are reluctant to release their logs for research purposes following the AOL incident in which personal information about people was accidentally put in the public domain (AOL, 2010). However, this can be overcome by projects such as the Lemur Query Log Toolbar which allows users intentionally to have their queries logged (Lemur, 2010).

4. QA & QLA - Common Areas and Research Questions

In this section we try to state briefly what we consider to be the key areas which are common to both, the principal research questions in these areas, and the means by which they might be investigated.

First of all, how real are the questions in QA and the queries in QLA? In QA, the questions are not usually from real users, they are devised by the assessors at CLEF, TREC etc. Secondly, they are restricted to certain well-known simple types which are only a small subset of the real questions which people wish to ask. These simplifications are necessary due to the limitations of our present day QA systems. Thirdly, questions are considered in isolation (or in some tracks a fixed group) and not in a dialogue context whereas in our interactions with people all questions are answered in context and with the possibility for clarification (see however Webb and
In this final section, we outline the main contributions of the papers accepted for the workshop and we attempt to link them together. These contributions address some of the research questions posed above.

Bernardi and Kirschner - From artificial questions to real user interaction logs: Real challenges for Interactive Question Answering systems.

This paper focuses on the issue of real logs vs. not real QA questions at TREC etc. There are three question sets: TREC, Bertomeu (collected in a controlled Wizard-of-Oz study) and BoB (a chatbot working at a university library site). These are analysed in respect of several different measures comparing utterances in a QA dialogue. The main conclusion is that the TREC data differs significantly from BoB in important respects such as length of query (BoB queries are shorter) and number of anaphora (Bob queries have less). The thinking is that in future TREC-style evaluations, questions should take into account these factors to make them as realistic as possible.


This paper compares queries submitted to a web search engine, queries submitted to a Q&A service (answers.com), and those used at TREC and CLEF in the QA tracks - six collections in all. This is very interesting because it is a direct comparison between the QA side and the QLA side. The core of the paper deals with an experiment in which well formed questions from answers.com are converted into IR-style queries (e.g. using just content words) and then a naïve Bayes classifier is used to try to recover the expected answer type and the original wh-word frame. For example "capital Ethiopia" should become "What is the capital of Ethiopia?" and the answer type is capital city. The thinking behind this interesting study is that if log queries can be converted to questions they can be answered exactly by a QA system.

Zhu et al. - Question Answering System Based on Community QA.

This paper considers whether sites such as Yahoo Answers - which contain millions of submitted questions and answers to them - can be used as a log-like resource to improve question answering. Given and input query, similar queries are then identified in the logs and their answers retrieved. A summarisation algorithm is then used to select sentences from the retrieved answers which can be used as the response to the input query.

Momtazi and Klakow - Yahoo! Answers for Sentence Retrieval in Question Answering.

This paper is also concerned with Yahoo Answers and its potential for improving QA performance. The authors developed two statistical frameworks for capturing relationships between words in a question-answer pair within Yahoo Answers. These were then used in a sentence selection task using as input TREC 2006 queries. Their best results exceeded the baseline which was considered to be the word based unigram model with maximum likelihood estimation.

Small and Strzalkowski - (Tacitly) Collaborative Question Answering Utilizing Web Trails.

The aim of the work is to study logs made by monitoring users in an interactive QA study. The information saved includes the question answered, the responses given and
those documents actually saved by each participant. Documents saved are placed in a standard order to allow comparisons between different searchers working on the same task. The key result is that in a study of 95 episodes, there is quite a degree of overlap between sets of files saved. This suggests an opportunity for sharing of data. One possible means of doing this is to observe a sequence of documents saved by a user, and when it overlaps with a previously observed sequence produced by another user, to offer the remainder of that saved sequence. This paper is interesting because it is the only one which collects QA data in a naturalistic setting, albeit within a controlled experiment where users are given predetermined tasks.

Sutcliffe, White and Kruschwitz - Named Entity Recognition in an Intranet Query Log.

This paper is concerned with queries in a highly focused log of searches conducted at a university web site. The authors firstly conducted a manual study some queries, categorising each by topic. In the process, a list of important named entity types was created. Secondly, training data for each NE type was created from the university website and this was used to train a maximum entropy NE tagger on a much larger log. This was evaluated, and statistics concerning NE occurrences in the log as a whole were computed. Finally, the possible use of NE data in answering the queries is discussed.

Mandl and Schulz - Log-Based Evaluation Resources for Question Answering.

This paper is concerned with the relationship between query logs and well-formed questions as answered by QA systems. The authors propose a system which can switch between IR-mode and QA-mode, depending on the input. They first discuss some of the log resources which are available for this kind of work, together with related Log analysis tracks at CLEF, and then present a preliminary analysis of question-like queries in the MSN log.

6. Conclusion

QLA and QA appear to be fields which intersect in a number of ways which suggest new research challenges for both. We listed some of these in Section 4. The papers accepted for the workshop address some of these challenges but not all of them. The paper by Bernardi and Kirschner compares real and created questions and shows how users really behave in QA systems as opposed to how evaluation designers think they behave (research question 6). A related paper by Leveling also compares the two kinds of questions. In addition, it tries to learn how the information need can be deduced from a short IR query (research question 1). However, a related aim, to use logs to develop better evaluation sets (research question 3) has not been addressed in depth.

The study of Momtazi and Klakow takes a deeper look at a real world QA service and its logs (research question 5). The same goal is pursued by Zhu at al. but they apply a different algorithm. A user study concerning a complex information need is presented by Small and Strzałkowski. Theirs is the only approach which addresses the issue of user sessions (research question 4). Sutcliffe at al. show for a real log file how NLP techniques can support the analysis (research question 2). Finally, research question 3 is again addressed by a short paper of Mandl and Schulz which argues that current logs for IR evaluation can also be a source of QA style questions.

The issues of click-through data and spoken language QA (research questions 7 and 8) have not been addressed at this workshop. Indeed, the other research questions have not generally been addressed in depth. However, several contributions used real-world QA resources and dealt with their properties in general and their differences from QA evaluation resources.

In summary, the papers provide an interesting survey of work in this developing field. However, while much has been achieved, all the contributors suggest interesting and worthwhile avenues for further research linking QLA and QA.

7. References


