# Developing a Test-Bed Framework for Modelling, Indexing and Retrieving Information on the Web: The Terrier Platform (GR/R90543/01)

**Iadh Ounis**
Department of Computing Science
University of Glasgow
e-mail:ounis@dcs.gla.ac.uk

Project Web page: http://ir.dcs.gla.ac.uk/terrier

I believe that this cost-effective project has been very successful both in terms of its theoretical and practical outcome. Major achievements include: new parameter-free probabilistic Web information retrieval models, use of novel combination of evidence approaches based on information theory to optimise retrieval on large-scale and heterogeneous document collections, development of a theoretically-founded hyperlink analysis approach for the Web, development of a state-of-the-art platform allowing the rapid development of large-scale information retrieval applications. 24 journal and conference papers were published as a result of this project, and an open-source version of the developed platform is currently distributed under an MPL license[1] for the benefit of the public in general and the information retrieval community in particular.

## 1 Background/context

Web Information Retrieval (IR) differs from classical IR in several aspects. An unprecedented amount of information is available on the Web, requiring effective and efficient search engines, which should handle a wide range of search requests from users. In addition, Web IR can exploit a number of different retrieval approaches, taking into account various sources of evidence, in addition to the textual content of documents. For example, in order to enhance retrieval effectiveness, we can use evidence from the document structure, from the hyperlink structure of the Web, or from the way documents are organised in a given site. However, the use of combination of evidence approaches that are parameter-free and adaptable to the document collections and the users' queries is remarkably rare, due probably to a lack of understanding of the characteristics of the Web and its users. Indeed, most of the combination of evidence approaches in the literature are rather ad-hoc, or based on weakly integrated search methods that are usually applied uniformly to all types of queries. However, not all queries benefit from the same sources of evidence, or from the same retrieval approaches. For example, the hyperlink structure of the Web does not improve the retrieval performance for precise information queries, due to the weaker nature of evidence obtained from the hyperlinks. Moreover, automatic query expansion/re-formulation retrieval approaches, while usually efficient for ad-hoc retrieval, are detrimental to the retrieval effectiveness when the first pass retrieval is of poor quality. Commercial search engines such as Google might have a good understanding of the Web, its characteristics and its users; however they do not usually share their findings with the scientific community.

Since 1999, the Text REtrieval Conference (TREC), an internationally acclaimed forum organised by the National Institute of Standards and Technology (NIST) in the USA, has dedicated a special track for the Web, supporting and encouraging research in this field. The main objective of the Web track is to assess the differences between Web and ad-hoc retrieval, improving our understanding of the search process on large-scale Web document collections. The Web track was indeed designed to allow a realistic and large-scale laboratory testing of Web IR techniques.

In this project, in collaboration with Dr. Gianni Amati from the Fondazione Ugo Bordoni, Italy, our aim was to develop nonparametric probabilistic Web IR models which are not only theoretically-founded, but can also automatically learn from empirical data and adapt to the users' information needs and queries. By developing a new IR platform called Terrier (Terabyte Retriever), and using it in the context of the TREC forum, we also aimed to demonstrate that the proposed approaches can deal with large-scale data collections, a necessary requirement in the context of Web IR, and operate in both a centralised and distributed setting effectively and efficiently.

[1] http://www.mozilla.org/MPL/

## 2 Key advances and supporting methodology

The practical aim of this project was to apply advanced probabilistic retrieval approaches, which combine efficiently and effectively various sources of evidence to improve the retrieval effectiveness on the Web. The research involved a number of strands, as follows.

### 2.1 Development and optimisation of parameter-free probabilistic models for IR

We used the Divergence from Randomness (DFR) paradigm described in Amati and Van Rijsbergen (2002), a generalisation of one of the very first models of IR, Harter's 2-Poisson indexing-model. The DFR framework is developed around a very elegant idea: the more the divergence of the within-document term-frequency of a term t from its distribution within the collection, the more the amount of information carried by t in the document. The DFR framework allows the generation of various statistically different document weighting models for document ranking, and term weighting models for query expansion.

The use of parameter-free probabilistic models is new in the IR field and is crucial, when dealing with large-scale collections of documents such as the Web. Indeed, it avoids the constraints of using costly parameter tuning approaches based on relevance assessment. During this project, we have refined and optimised the DFR framework, so that it automatically learns from the document collection, and adapts to it as well as to the users' queries. We have developed a new and theoretically-driven term frequency normalisation methodology, described in He and Ounis (2003), which was shown to work extremely well on various large-scale collections and retrieval tasks. For example, using our proposed methodology, we achieved the best retrieval performance in the TREC 2004 Terabyte track, where experiments were conducted on the largest ever IR document test collection, built from a large crawl of the Web. Moreover, the proposed methodology has allowed us to achieve an outstanding retrieval performance in the CLEF 2004 French monolingual retrieval track, and the TREC 2003 and 2004 Robust tracks.

We have also proposed a novel weighting function recommender mechanism, initially described in He and Ounis (2004a), that involves a query clustering process allocating the optimal DFR-based weighting models, including document ranking and query expansion models, on a per-query basis. The approach was shown to be very effective and efficient in the TREC 2003 and 2004 Robust tracks.

### 2.2 Combination of evidence and selection of retrieval approaches on the Web

Most existing approaches for Web IR have used a uniform retrieval approach for each query. We have shown in Plachouras and Ounis (2004c) and Plachouras *et al.* (2005a) that not all queries benefit equally from the same retrieval approach, especially in the Web context, given the diversity of retrieval tasks undertaken by users. Therefore, we have developed an original and extensible Bayesian decision mechanism, which employs evidence from the textual content, URL type and the hyperlink structure of the set of retrieved documents to optimally select the appropriate retrieval approach on a per-query basis.

This new decision model uses a precursor approach based on Information Theory and various statistical measures such as query scope, distribution of query terms in anchor texts, and domain aggregates (Plachouras and Ounis (2004b)) to assess and predict the usefulness of a particular retrieval approach. We have also shown how the parameters of the decision mechanism can be automatically and efficiently learnt from the document collection, without the need of relevance assessment. The proposed new approach has a very marginal overhead and has been demonstrated to be extremely efficient and effective on various large-scale Web collections, as demonstrated by our outstanding participation in the TREC 2002, 2003 and 2004 Web tracks.

As the project progressed, we investigated other novel research topics, such as the dynamic selection of retrieval approaches, based on the predicted performance of a query. Indeed, we have developed state-of-the-art query performance pre-retrieval predictors, described in He and Ounis (2004b). These predictors estimate the performance of a given query before the retrieval process takes place. For instance, approaches such as query expansion/re-formulation techniques are not considered when the query is predicted to generate a poor first pass retrieval performance. All proposed predictors have a very marginal overhead, and were shown to have a very good performance in the TREC 2004 robust track.

### 2.3 Citation/Web Hyperlink Analysis

Most current commercial search engines incorporate a link analysis component in their document ranking mechanism. For example, Google's PageRank algorithm or IBM's HITS approach use link analysis to discover documents of high quality on the Web. However, these approaches are rather ad-hoc, and characterised as being heuristics.

During this project, we have developed the *first* theoretically-founded link analysis approach, called the Absorbing Model, described in Plachouras et al. (2005b), which is based on Markov Chains, and which has been successfully used in our large-scale TREC 2002 and 2003 Web IR experiments.

The Absorbing Model has a similar computational cost to other existing approaches, and is parameter-free, with no use of parameters such as the PageRank's damping factor. Moreover, unlike the existing link analysis approaches, it can be applied in both a query-dependent and query-independent manner. The generality of our approach is such that it is opening up a wide range of novel applications (e.g. multilingual retrieval). We are currently investigating a patent application.

### 2.4 Development of the Terrier information retrieval platform:

Experience has shown that having a system allowing large-scale experimentation to be conducted in a robust, transparent, reproducible, modular and platform-independent way, without constraints, is very important and crucial in IR. This is demonstrated by the fact that the major leading IR groups in the World have all a proprietary system: e.g. CMU has Lemur, UMass has Inquery, CSIRO has Panoptic, RMIT has Zettair, and Microsoft has Okapi.

During this project, we have built the original Terrier (Terabyte Retriever) platform, described in Ounis *et al.* (2005), which combines new and cutting-edge ideas from probabilistic theory, hyperlink analysis, automatic query expansion/re-formulation methodologies, and data compression techniques. Currently, the Terrier platform deploys more than 50 Divergence From Randomness (DFR) models for document ranking. It is also able to predict query performance, and automatically select the optimal document weighting model and/or the appropriate retrieval approaches (e.g. query expansion, anchor text information, or link analysis) on a per-query basis. Terrier has significantly boosted our participation in the internationally acclaimed TREC forum since 2002.

In addition to being a vehicle for the evaluation of all new approaches proposed within this project, Terrier allowed us to easily simulate, assess and improve state-of-the-art IR technology, in particular on Web collections. Terrier enabled us to investigate state-of-the-art efficient storage and compression techniques, based on gamma and unary encodings, which allowed a cost-effective indexing of large document collections such as the Web. For example, a full-text index of the TREC Terabyte track .GOV2 collection, the total size of which is 426GB, takes 17.48GB, including the inverted indices (7.77GB), the direct indices (7.00GB), the lexicons (1.84GB) and the document indices (0.87GB). This corresponds to only 4.1% of the total collection size. As demonstrated by the TREC 2004 Terabyte track results, this is by far the best compression rate currently available in the IR community.

Moreover, as described in Cacheda et al. (2005), we have also investigated different architectures for a distributed IR system, a necessary setting to deal with large-scale collections such as the Web. Taking into account the available infrastructure resources, this study allowed us to use the optimal system architecture in the TREC 2004 Terabyte track, as well as the intranet search facilities we are currently deploying for the Department of Computing Science at the University of Glasgow, and many other organisations. The importance of the Terrier platform is twofold:

(1) Terrier allowed us to improve the current understanding of theoretical IR, especially the behaviour of search engines on very large-scale collections
(2) It is a test-bed framework allowing for the rapid experimentation of new IR concepts/ideas, which boosted our research in Web IR and our participation in TREC 2002, 2003 and 2004, where our results were outstanding. In particular, in 2004, we participated in 3 tracks of TREC: Robust track, Web track and Terabyte track.

## 3 Project plan review

The project went as planned. All research conducted within this project was fed into the Terrier software, resulting in a cutting-edge state-of-the-art IR test-bed platform.

As initially expected, the project served as a training framework and a development platform for our undergraduate and postgraduate students. As a consequence, it has attracted and significantly benefited from the important contribution of many students. Indeed, the students' input greatly helped in speeding up the project progress, and contributed to its success.

Vassilis Plachouras, a computing science research student, made a significant contribution to the project by researching various new Web IR approaches, especially when and how to use the Web linkage information, and how to integrate it with other sources of evidence. He developed various modules of the Terrier platform, and contributed significantly to its overall design and architecture.

Ben He, a computing science research student, contributed actively to the project, especially to the querying part of the Terrier platform. He developed various important modules of the software, including many optimisation techniques, and significantly contributed in evaluating the Terrier platform.

Craig Macdonald, initially a computing science undergraduate then a research student at the Department, focused on developing a crawler for the Terrier search engines, called Labrador, allowing various intranet search facilities to be deployed. His software engineering skills greatly helped in the refinement and improvement of the overall design and architecture of Terrier, which can be now easily used for a wide range of applications, such as an experimental IR test-bed, a search engine, or a desktop search facility.

Many research students in the Department are currently using Terrier as a development platform. For example, Christina Lioma uses Terrier for multilingual retrieval, and successfully participated in CLEF 2004. Jianqiao Liu uses Terrier for XML retrieval, and Azreen Azman uses Terrier for various users profiling activities, allowing a personalised retrieval.

In addition, many undergraduate students have actively participated to the project. Douglas Johnson has developed the flexible proximity search functionality of Terrier, and integrated it to the indexing component of the system. Sau Kwan Chan has developed various string matching algorithms for Terrier, allowing for automatic spelling errors detection and correction. Rachel Tsz-Wai Lo developed an automatic DFR-based approach allowing the extraction of stopwords from a given collection. Other undergraduate and postgraduate student projects are still ongoing, using Terrier as a development platform.

The project has also benefited from a very successful interaction with Dr Fidel Cacheda at University of A Coruña, Spain, on the use of distributed architectures in building large-scale IR systems. Some of the work done in collaboration with Dr Cacheda has been used in the design of the distributed architecture of Terrier, which enabled us to handle very large test collections.

## 4 Research impact and benefits to society

The publication of various Web retrieval approaches has exposed researchers to new ways of combining evidence on the Web. Our work has yielded significant insights into the behavior of IR techniques on very large-scale collections, and allowed to create a state-of-the-art technology and a significant test-bed platform for IR research.

We have been invited to give a plenary presentation on our research work at the highly influential International Text REtrieval (TREC) Conference, in three successive years: November 2002, November 2003 and November 2004. In November 2002 and 2003, we were invited to the Web track session, while in November 2004, we were invited to the Robust track session. Invitations to present plenary papers at TREC are based on originality of research, resulting system's performance and potential industrial impact. In particular, in November 2004, our Terrier platform has achieved the best official overall run in the Terabyte track, where experiments were conducted on the largest ever IR test collection. Moreover, Terrier achieved the best official topic distillation run of the TREC 2004 Web track. Overall, our Web search models can be seen as state-of-the-art approaches, as asserted by the TREC conference

results, and the various publications in the reputable IR conferences and journals: e.g. SIGIR, CIKM, RIAO, SPIRE, ECIR, WWW, Journal of Information Retrieval, Journal of Web Engineering, and Journal of Information Processing and Management.

For the benefit of the public, a version of Terrier is now available as open source software, which is distributed under the Mozilla Public License (MPL). The aim is to facilitate experimentation and research in IR. Since the very recent release of the software, over 40 well-established academic and industrial institutions have expressed interest and downloaded the software, some of which are affiliated to major institutions around the World. These include University of Padova, Italy, QMU, London, Ajou University, South Korea, University of California, USA, University of Sao Paulo, Brazil, University of Tampere, Finland, University of Maryland, USA, University of Santiago de Compostela, Spain, the Chinese Academy of Science, Toyohashi University of Technology, Japan, the Rochester Institute of Technology, Swedish School of Library and Information Science, Royal School of Library and Information Science, Denmark, Haifa University, Israel, and RMIT, Melbourne.

Several public organisations in Britain, in Italy and in the USA have expressed interest in using Terrier for their intranet search. In particular, the British Computing Society (BCS) and the Italian Ministry of Communications are currently deploying Terrier for their in-house intranet search. Moreover, corporations like Torg Consulting, Fujitsu and Memex have also shown interest in using the system. In addition, with the help of our colleague Prof. Theo Huibers, director at KPMG and professor at the University of Twente, Netherlands, many important public organisations in the Netherlands are considering to use Terrier as their intranet search engine.

Furthermore, Terrier is currently used to search the intranet of the Department of Computing Science at the University of Glasgow, and is currently being deployed for the intranet of the University of Glasgow domain and the digital libraries of the HATII Department at the University of Glasgow (further information can be found in http://ir.dcs.gla.ac.uk/terrier/applications.html).

Finally, Terrier is currently the main development platform for both undergraduate and postgraduate students in our research group, allowing them to employ a state-of-the-art framework for their research. In particular, various and diverse projects are currently developed on top of the Terrier framework. Moreover, the University of Strathclyde, UK, is currently using it for teaching purposes.

## 5 Explanation of expenditure

Expenditure was much as expected. We allocated money to visit the Fondazione Ugo Bordoni, Rome, to discuss the research progress/outcome as well as the progress of the platform. We also used the money for travel to international conferences, in order to disseminate the abundant results and research papers generated within this project. In particular, we travelled to the TREC International Conference, ACM SIGIR & ACM CIKM Conferences, ECIR, WWW, RIAO, and SPIRE.

## 6 Further research or dissemination activities

As planned in the original proposal, we presented the research at a range of reputable conferences, so that to disseminate results in various important forums, e.g. ACM SIGIR 2004, ACM CIKM 2003, ECIR 2005, ECIR 2004, RIAO 2004, SPIRE 2004, WWW 2004, etc. We have also presented the research at the International TREC conference in all successive years. Papers have been also submitted to a diverse range of journals, e.g. Journal of Information Retrieval, Journal of Web Engineering, Journal on Digital Information Management, and Journal of Information Processing and Management (see http://ir.dcs.gla.ac.uk/terrier/publications.html for further details). 24 papers have been published so far as a result of this project, and still are to come. Indeed, as well as the papers already published, some papers are currently awaiting results for acceptance (e.g. paper submitted to the Journal of Information Systems), and some papers are in preparation to be submitted to ACM SIGIR 2005 and to the Journal of Information Processing and Management.

Many public organisations are more and more interested in the Terrier framework. The Dutch Public Libraries organisation (http://www.bibliotheek.nl) has expressed an interest in the software for its intranet. We will be demonstrating the system to them in February. In case of an agreement, then every Dutch public library internet site will use Terrier.

We will refine some research aspects of the project in some follow-on projects: Leverhulme Trust project on length normalisation in IR, and Carnegie Trust project on deploying Terrier in a domain-dependent application, namely on a large-scale bioinformatics collection.

Furthermore, various cutting-edge modules have been developed in the context of the Terrier project. For example, a novel link-analysis module that is an alternative to Google's PageRank, and which we might consider for a patent because of its various potential applications. Our medium-term goal is to build a British/European search engine (the first one actually), which will cover all British/European-based Web pages, resulting in an even larger document collection, where the full potential of the Terrier's technology can be evaluated and refined. Our ultimate goal is to run a search engine on the full Web (or as much as we can accommodate using a particular infrastructure). Indeed, we believe that Terrier implements the state-of-the-art on Web search technology and has all the prerequisites to become the European answer to Google. We intend to apply for an EPSRC follow-on project.

## 7 Conclusions

I believe that this has been a very successful and cost-effective project, which has led to a state-of-the-art IR test-bed framework, as well as many novel probabilistic Web retrieval models. The research has already led to a significant number of generated and published results, many of which were not foreseen at the start of the project. The collaboration expanded to include a number of international colleagues from Spain and the Netherlands, and will continue in follow-on projects.

## References


Amati,G., van Rijsbergen,C.J. (2002). Probabilistic models of information retrieval based on measuring divergence from randomness. ACM Transactions on Information Systems, 20(4):357-389.

Cacheda,F., Plachouras,V., Ounis,I. (2005). A case study of distributed information retrieval architectures to index one terabyte of text. Information Processing and Management, Elsevier Science, In Press.

He,B., Ounis,I. (2003). A study of parameter tuning for term frequency normalization. Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM), New Orleans, USA, ACM, pp 10-16.

He,B., Ounis,I. (2004a). A query-based pre-retrieval model selection approach to information retrieval. Proceedings of the 8th International Computer Assisted Information Retrieval (RIAO) Conference, Avignon, France, CID Press, pp 706-719.

He,B., Ounis,I. (2004b). Inferring Query Performance Using Pre-retrieval Predictors. Proceedings of 11th Symposium on String Processing and Information Retrieval (SPIRE), LNCS 3246, Padova, Italy, Springer, pp 706-719.

Ounis,I., Amati,G., Plachouras,V., He,B., Macdonald,C., Johnson,D. (2005). Terrier Information Retrieval Platform. Proceedings of the 27th European Conference on Information Retrieval (ECIR), Springer, To Appear.

Plachouras,V., Cacheda,F., Ounis,I., (2005a). A decision mechanism for the selective combination of evidence in topic distillation. Information Retrieval, In Press, Kluwer Academic Publisher.

Plachouras,V., Ounis,I., Amati,G. (2005b). The Static Absorbing Model for Hyperlink Analysis on the Web. Journal of Web Engineering, Rinton Press, Princeton, In Press.

Plachouras,V., Ounis,I., Cacheda,F. (2004a). Selective Combination of Evidence for Topic Distillation using Document and Aggregate-level Information. Proceedings of the 8th International Computer Assisted Information Retrieval (RIAO) Conference, Avignon, France, CID Press, pp 610-622.

Plachouras,V., Ounis,I. (2004b). Distribution of Relevant Documents in Domain-level Aggregates for Topic Distillation. Proceedings of the 13th International ACM World Wide Web Conference (WWW ), New York, USA, pp 372-373, ACM Press.

Plachouras,V., Ounis,I. (2004c). Usefulness of hyperlink structure for query-biased topic distillation. Proceedings of the 27th ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, ACM Press, pp 448-455.