Report on the Scottish Informatics and Computing Science Alliance's Information Retrieval Workshop (IRFest)

Martin Halvey
Glasgow Caledonian University
martin.halvey@gcu.ac.uk

Leif Azzopardi
University of Glasgow
Leif.Azzopardi@glasgow.ac.uk

Abstract

The 1st Scottish Information Retrieval Workshop took place in May, 2013 in Glasgow. The focus of the workshop was on bringing together IR researchers from the various Scottish universities in order to facilitate more awareness and increased interaction. The scientific program included a minute madness session, research talks, demos and posters. The keynote was delivered by Prof. Arjen de Vries who provided an overview of the problems, opportunities and challenges of finding strategic information within the enterprise.

1 Overview

On the 31st of May, 2013, the 1st SICSA Scottish Information Retrieval (IR) Workshop was held at Glasgow Caledonian University. The workshop was organised by the Glasgow Caledonian University and the University of Glasgow, and was funded by the Scottish Informatics and Computing Science Alliance (SISCA). The aim of the workshop was to increase awareness and interaction between IR-related researchers working in, across, and with Scottish universities.

The workshop had a total of 41 SICSA attendees, comprised of 30 registered attendees and 11 invited speakers from 8 different SICSA institutions (Glasgow Caledonian University, University of Glasgow, Strathclyde University, University of Aberdeen, Robert Gordon University, Heriot Watt University, University of Edinburgh, and University of Dundee Abertay). The keynote speaker was Prof. Arjen de Vries from CWI, Amsterdam. Prof. Keith van Risjerbergen and Dr Robert Villa also made special guest appearances.

The scientific program included a keynote talk, a minute madness session, research talks, demos and posters. The madness session allowed attendees to outline their research interests providing an excellent overview of the novelty and diversity of IR research being undertaken in Scotland. These interests included (but were not limited to): personal music retrieval, social media filtering for emergency management, retrieval of cultural and heritage objects, ranking for aggregated search, developing models of findability, processing social media for crime analysis, compression and efficiency, and using context for app recommendations. After the introductions, the keynote was delivered by Prof. Arjen de Vries (CWI, Amsterdam), who provided an excellent overview of the problems, opportunities and challenges of finding strategic information within enterprise.

The SICSA speakers then covered a wide range of topics. Jesus Rodriguez Perez, (University of Glasgow) talked about using inter-document relations in microblog retrieval. Prof. Miles Osborne (University of Edinburgh) discussed efficient cross stream event detection social media, in particular for event detection in twitter. Dr Tiphaine Dalmas, (University of Edinburgh) described question answering with Spacebook, which is a speech-driven, hands-free, eyes-free mobile application for pedestrian navigation and exploration in urban environments. Dr Matt-Mouley Bouamrane (University of Glasgow) discussed information management in the patient surgical pathway in NHS Scotland. Dr Nava Tintarev (University of Aberdeen) presented work from SaSSY which examines explanations in scrutable autonomous systems (agents, planning and argumentation dialog). Prof. Ayse Goker (Robert Gordon University) discussed the benefits of adopting a user centred approach for designing multimedia IR systems. Dr Craig MacDonald (University of Glasgow) talked about sensor and social search within smart cities. Following on from this Dr M-Dyaa Albakour (University of Glasgow) described his approach to local event retrieval with social sensors. Dr Martin Halvey (Glasgow Caledonian University) outlined an examination the effort involved in making relevance assessments as part of the information retrieval process. Dr Dmitri Roussinov (University of Strathclyde) outlined the use of web n-grams to automatically recognize important aspects of products in opinions when shopping. Dr Leif Azzopardi (University of Glasgow) outlined how economics can be used to model the interaction between a user and system in the context of search. The event concluded with a discussion session which centred on future SICSA IR events and collaboration amongst the community.

2 Summary of Talks

Keynote: Looking beyond plain text for document representation in the enterprise! by Prof. Arjen de Vries. In his keynote, Arjen highlighted that in today's academic institutions, strategic questions are those that relate to dependency on funding instruments, the public private partnerships that exist, the match between topic areas addressed by the research staff and those claimed important by policy makers. The professional search tasks encountered to answer questions in this domain are usually addressed by business intelligence (BI) tools, and not by search engines. However, professionals are known to be busy people inspired by their own research interests, and not particularly fond of keeping the customer relationship management (CRM) or knowledge management systems up to date for the organisation's strategic interest. Instead of requiring research staff (or their administrative support) to provide this management information, Arjen illustrated how the desired information usually exists already in the documents inherent to the academic work process. Therefore, he argued, that information retrieval could play an important role in the computer systems that support the business analytics involved, and could significantly improve the coverage of entities of interest - i.e. reducing the effort involved in achieving good recall in business analytics. The ranking functionality over the enterprise's (textual) content should however not be an isolated component. In the context of academia, Arjen explained how the information derived from research proposal, research publications and the financial systems could be integrated providing an excellent motivation and opportunity for a more unified approach to structured and unstructured data.

On Using Inter-Document Relations in Microblog Retrieval by Jesus Rodriguez Perez, University of Glasgow. In his talk, Jesus described the problem of vocabulary mismatch in the context of Twitter, and how in this domain the high diversity of language and sparsity of the data exacerbates the mismatch between the query and relevant tweets. To overcome this problem he presented a re-ranking approach relying on inter-document relations, which attempts to bridge this gap. Experiments with TRECs Microblog 2012 collection showed that including such information in the retrieval process improved retrieval effectiveness - providing him with a strong baseline to further improve upon [16].

Cross stream event detection by Prof. Miles Osborne, University of Edinburgh. Following on from the previous Twitter related talk, Miles described his research which focused on whether social media could be used as a source of real-time breaking news [15]. For example, when Osama Bin Laden was killed by US forces the news was first made public on Twitter. Miles argued that rapidly finding all breaking news has clear economic and humanitarian benefits, but finding all such breaking news presents hard computational challenges. Specifically, he pointed out that to detect news-related novelty in massive streams (upwards of two thousand posts per second) needs to be performed as quickly as possible. Efficiency is not the only consideration, and that the problem of dealing with enormous quantities of irrelevant posts also needs to be handled. He outlined how he tackled the first problem using Locality Sensitive Hashing, taking constant time per post. And then, how he used Storm to parallelise this computation, yielding a system capable of processing 2k tweets per second. The second problem was tackled by intersecting the Twitter stream with Wikipedia page requests, to filter-out spurious first stories. Taken together, this resulted in processing more than 250 million items per day. More details on Miles work on this topic can be found in [12–14].

Question Answering for Spacebook by Dr Tiphaine Dalmas, University of Edinburgh. In her talk, Tiphaine described Spacebook; Spacebook is a EU-funded FP7 project based on EARS [6]; the system has been described in [10] and the preliminary evaluation has been presented at ACL 2013 ([11]). In her talk Tiphaine focused on describing the task of pedestrian exploration (i.e. a tourist wandering around a city) and described the question answering techniques used in combination with GIS technology in the dialogue system. A prototype system was developed for the city of Edinburgh, and Tiphaine reported on the preliminary evaluation that was performed on the streets with tourists.

A study of Information Management in the Patient Surgical Pathway in NHS Scotland by Dr Matt-Mouley Bouamrane, University of Glasgow. In his talk, Matt described a study of information management processes across the patient surgical pathway in National Health Service (NHS) Scotland [7–9]. While the majority of General Practitioners (GPs) consider electronic information systems as an essential and integral part of their work during the patient consultation, Matt explained that many were not fully satisfied with the functionalities of these systems, and that the quality of discharge information varied widely across the nation. Matt concluded that there was insufficient use made of information provided through the patient electronic referral and there was a considerable duplication of effort with the work already performed in primary care. However, in the three health-boards that have implemented electronic preoperative information systems their clinical practices have been transformed, facilitating better communication and improved information sharing amount multidisciplinary teams. Finally, he concluded that the next major challenge is ensuring that the right information, the right amount of information is delivered to the right person as a patient interacts with the NHS.

SaSSY by Dr Nava Tintarev, University of Aberdeen. An autonomous system consists of physical or virtual systems that can perform tasks without continuous human guidance. These types of systems are becoming increasingly ubiquitous, ranging from unmanned vehicles, to robotic surgery devices, to virtual agents which collate and process information on the internet. Existing autonomous systems are opaque, limiting their usefulness in many situations. With this in mind Nava presented work from the Scrutable Autonomous Systems (`http://www.scrutable-systems.org/`) project which aims to enable scrutiny of autonomous systems [17]. Nava presented a prototype system for presenting plans in a human readable form. In addition, she outlined three initial evaluations into plan presentation, the first looked at labelling hierarchical tasks, the second syntactic aggregation and the third vague expressions. These evaluations are ongoing and it is hoped that they will lead to a better understanding of information presentation for autonomous systems.

Challenges in multimedia information retrieval from a user-centred perspective by Prof. Ayse Goker, Robert Gordon University. In her talk, Ayse presented an excellent overview of the different problems one faces when evaluating IR systems with users. While, multimedia content is increasingly abundant, accessible, and vast, Ayse pointed out that our means of presenting results to users are still primarily based on the text paradigm. She argued that this is insufficient and often unsuitable for users, particularly those who deal with images and videos regularly for their professional tasks such as those in the creative industries and journalists. Using examples from several projects, including EU Social Sensor, Ayse highlighted the emerging challenges in multimedia information retrieval.

Sensor and Social Search within Smart Cities by Dr Craig MacDonald, University of Glasgow. In his talk, Craig discussed issues to do with local (concerned with entities close by), timely (concerned with events happening right now) search. His talk presented an overview of challenges and interaction paradigms that could drive future IR research, covering IR architecture and modelling, as well as the evaluation methodologies and datasets needed to conduct this research. Craig envisaged that many local, timely information needs could be addressed by fusing information gleaned from sensors with the increasingly ubiquitous social networks and that this type of infrastructure could be provided in future smart cities. With this in mind he illustrated ongoing work within the related SMART FP7 project which addresses search scenarios within smart cities. (for more details see[1, 2] and `http://www.smartfp7.eu/`)

Local Event Retrieval with Social Sensors by Dr M-Dyaa Albakour, University of Glasgow. Following on from the previous talk on local search, M-Dyaa presented a new event retrieval framework to locate an event happening in a certain area within a city that matches a user generated query[1, 2]. The framework measures unusual microblogging activities in a certain area and uses that as an indication of the occurrence of an event. As well as presenting a new retrieval framework, M-Dyaa also presented a novel evaluation methodology for local search that is inspired by the conceptually similar IR problem of video segmentation. Using this methodology the framework was evaluated using a set of tweets collected over a period of twelve days from different areas of London, as well as two sets of local events collected within the same period using crowdsourcing and local news sources in London. The results show that the proposed event retrieval framework is capable of identifying and ranking events within a city. However, when applied on multiple fine-grained areas within the city, the retrieval effectiveness degrades, this was due to the nature of the events considered in the experiments, i.e. their low coverage on Twitter. Research is ongoing to address the caveats observed in the evaluation.

Is relevance hard work? Evaluating the effort of making relevant assessments by Dr Martin Halvey, Glasgow Caledonian University. Martin outlined a user evaluation which measured the effort users must exert to judge the relevance of document, investigating the effect of relevance level and document size. While the criteria by which assessors judge relevance has been intensively studied, little work has investigated the process individual assessors go through to judge the relevance of a document. Martin argued that by better understanding the process and effort involved in making relevance judgements, we may provide data which can be used to create better models of search. Results from initial evaluations suggest that relevant documents require more effort to judge when compared to highly relevant and not relevant documents, and that effort increases as document size increases [18]. The talk concluded with a discussion of future directions for measuring effort involved in various part of the search process.

Using Web N-grams to Automatically Recognize Important Aspects of Products in Opinions by Dr Dmitri Roussinov, University of Strathclyde. Dmitri presented approach for mining buyers opinions to automatically detect which aspects of products, e.g. screen size and battery life for a mobile phone, are most important. His machine learning approach combines recognizing certain syntactic patterns with validating semantic relationships between the products and their aspects, such as part of, result of, characteristic of, etc. Dmitri validated this approach through a statistical analysis of occurrence of certain patterns (e.g. camera features lcd screen, sound of ipod, etc.) in the entire Web corpus by involving Microsoft Bings N-grams service.

The Economics of Searching by Dr Leif Azzopardi, University of Glasgow. During his talk, Leif outlined how microeconomic theory, could be applied to model the interactive information retrieval process [4]. This provided a way to formally model the interaction between a user and a system. He argued that by such models make it is possible to: (1) theorise and predict how users will behave when interacting with systems, (2) ascertain how the cost and performance influences interaction, and (3) understand why particular interaction styles/strategies/techniques are/aren't adopted by users. Essentially, he argues that such economic models of the search process can provide explanations as to why we observe particular user behaviours[3–5].

3 Outlook

Overall the 1st SICSA Scottish Information Retrieval Workshop was a success creating new relationships and a greater awareness of the exciting and dynamic research being undertaken in Scotland. Tentative plans to organize the 2nd workshop next year were formed, along with plans to invite distinguish researchers to visit and tour Scotland.

4 Acknowledgements

We would like to thank the Scottish Informatics and Computing Science Alliance for their sponsorship, Glasgow Caledonian University for hosting the event and sponsoring the evening reception, and Prof. Arjen de Vries for delivering an interesting and engaging keynote.

References

[1] M.-D. Albakour, C. Macdonald, I. Ouis, A. Pnevmatikakis, and J. Soldatos. Smart: An open source framework for searching the physical world. In *Proceedings of the SIGIR Workshop in Open Source Information Retrieval*, 2012.

[2] M.-D. Albakour, C. Macdonald, and I. Ouis. Identifying local events by using microblogs as social sensors. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, Lisbon, Portugal, 2013. Springer.

[3] L. Azzopardi. Query side evaluation: an empirical analysis of effectiveness and effort. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 556–563, 2009.

[4] L. Azzopardi. The economics in interactive information retrieval. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 15–24, 2011.

[5] L. Azzopardi, D. Kelly, and K. Brennan. How cost affects search behaviour. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '13, pages 23–32, 2013.

[6] P. J. Bartie and W. A. Mackaness. Development of a speech based augmented reality system to support exploration of cityscape. *Transactions in GIS (special issue)*, 10(1): 63–86, 2006.

[7] M.-M. Bouamrane and F. Mair. An overview of electronic health systems development integration in scotland. In *Proceedings of the first international workshop on Managing interoperability and complexity in health systems*, MIXHS '11, pages 59–62, 2011.

[8] M.-M. Bouamrane and F. Mair. A study of general practitioners' perspectives on electronic medical records systems in nhsscotland. *BMC Medical Informatics and Decision Making*, 13(1):58, 2013.

[9] M.-M. Bouamrane, F. Mair, and C. Tao. Managing complexity in pre-operative information management systems. In *Proceedings of the first international workshop on Managing interoperability and complexity in health systems*, MIXHS '11, pages 3–10, 2011.

[10] S. Janarthanam, O. Lemon, X. Liu, P. Bartie, W. Mackaness, T. Dalmas, and J. Goetze. Integrating location, visibility, and question-answering in a spoken dialogue system for pedestrian city exploration. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 134–136, Seoul, South Korea, 2012.

[11] S. Janarthanam, P. Bartie, T. Dalmas, X. Liu, O. Lemon, B. Webber, and W. Mackaness. Evaluating a City Exploration Dialogue System Combining Question-Answering and Pedestrian Navigation. In *Proceedings of the Association for Computational Linguistics (ACL)*, 2013.

[12] M. Osborne, S. Petrovic, R. McCreadie, C. Macdonald, and I. Ounis. Bieber no more: First Story Detection using Twitter and Wikipedia. *SIGIR 2012 Workshop on Time-aware Information Access (#TAIA2012)*, 2012.

[13] S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 181–189, 2010.

[14] S. Petrović, M. Osborne, and V. Lavrenko. Using paraphrases for improving first story detection in news and twitter. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 338–346, 2012.

[15] S. Petrovic, M. Osborne, R. McCreadie, C. Macdonald, I. Ounis, and L. Shrimpton. Can twitter replace newswire for breaking news? In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[16] J. A. Rodriguez Perez, Y. Moshfeghi, and J. Joemon. On using inter-document relations for microblog retrieval. In *WWW 2013*, page p75. ACM, 2013.

[17] N. Tintarev, R. Kutlak, N. Oren, K. Van Deemter, M. Green, J. Masthoff, and W. Vasconcelos. Sassyscrutable autonomous systems. *Proceedings of The Society for the Study of Artificial Intelligence and the Simulation of Behaviour 2013*, 2013.

[18] R. Villa and M. Halvey. Is relevance hard work?: evaluating the effort of making relevant assessments. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '13, pages 765–768, 2013.