

ECIR Reviewing Workshop



BCS Information Retrieval Specialist group

20th June 2006

**Hosted by: The Department of Computing Science,
University of Glasgow**

**Organizers: Iadh Ounis, Leif Azzopardi, Andy MacFarlane
Editors: Leif Azzopardi, Andy MacFarlane, Iadh Ounis**

Table of Contents

	Page
1. Introduction	4
2. Questionnaire Responses and Feedback	6
3. Experiences from ECIR 2005 PC – contributed by David Losada	9
4. Meta-Reviewing – contributed by Ian Ruthven	10
5. Reviewing with Sub Programme Committees – contributed by Gianni Amati	12
6. How to write a good ECIR paper	14
6.1 Theoretical Papers – contributed by Keith van Rijsbergen	14
6.2 Experimental/comparisons Papers – contributed by Iadh Ounis	15
6.3 “People in IR” - User studies/Interfaces papers – contributed by Ian Ruthven	16
6.4 Applications and System Prototype papers – contributed by Ayse Goker	18
6.5 Conceptual – contributed by Mounia Lalmas, Leif Azzopardi	20
6.6 Evaluation and Measurement papers – to be added	20
6.7 Taxonomy of research papers as applied to IR – Adapted from Ian Parberry	20
7. Discussion of issues raised during the workshop	
Appendix A – Survey Questionnaire	

Organizers:

Iadh Ounis, University of Glasgow, Glasgow
Leif Azzopardi, University of Strathclyde, Glasgow (Chair)
Andy MacFarlane, City University, London

Attendees:

Gianni Amati, FUB, Rome
Ali Azim Bolourian, University of Glasgow, Glasgow
Francesca Crestani, University of Strathclyde, Glasgow
Ayse Goker, The Robert Gordon University, Aberdeen
Mounia Lalmas, Queen Mary, University of London
Monica Landoni, University of Strathclyde, Glasgow
Christina Lioma, University of Glasgow, Glasgow
David Losada, Universidad de Santiago de Compostela, Spain
Craig Macdonald, University of Glasgow, Glasgow
Michael Oakes, University of Sunderland Sunderland
Simon Overell, Imperial College, London
Vassilis Plachouras, University of Glasgow, Glasgow
Ian Ruthven, University of Strathclyde, Glasgow
Keith van Rijsbergen, University of Glasgow, Glasgow
Murat Yakici, University of Strathclyde, Glasgow

Acknowledgements:

The BCS-IRSG would like to thank the Department of Computing Science at the University of Glasgow for hosting the workshop. We would also like to thank all the workshop and survey participants for their contributions.

1. Introduction

A workshop on various issues to do with reviewing in ECIR (European Conference on Information Retrieval) was convened on 20th June 2006 at the University of Glasgow. The aims of the workshop were two-fold: (1) to review current practise of reviewing at ECIR, and provide a framework and guidance notes for future ECIR programme chairs; and (2) to develop a set of guidelines to help authors wishing to submit to ECIR and also to aid referees reviewing for ECIR.

Given the success of ECIR 2006 over previous instances of the conference (177 paper submissions to ECIR 2006 compared with 124 submissions a ECIR 2005), there was a need to rethink the management of the reviewing process, and a decision needed to be made on which method to use – status quo, meta-reviewing or Sub-PC's.

With the quantity of submissions also comes the problem of ensuring quality; in terms of the submissions themselves, the reviews, and ultimately the final programme. To ensure and maintain the high quality of the conference, it was felt that establishing a set of guidelines for ECIR would be helpful to both authors wishing to submit and referees reviewing papers. We dedicated a session on discussing what would make a good ECIR paper given some of the different genres.

A questionnaire was distributed beforehand to ascertain the general thoughts and feelings of the IR community on ECIR. The results were presented during the day and un-resolved issues discussed at the conclusion of the day.

The Workshop Agenda

Introduction –

Survey Results – presented by Leif Azzopardi

Session 1 - Structure/Organization of PC

Meta Reviewing – presented by Ian Ruthven

Sub PC Reviewing – presented by Gianni Amati

Session 2 - What makes a good IR paper?

Theoretical Papers – presented by Keith van Rijsbergen

Experimental – presented by Iadh Ounis

User studies – presented by Ian Ruthven

Applications – presented by Ayse Goker

Session 3 - Discussion

Rating papers, reviewing questions, other issues

This volume contains the result of the discussions and presentations of the workshop, and some further contributions subsequent to the workshop, which the editors felt necessary for completeness. Included in these proceedings are the results of the survey, overviews of the two main reviewing processes (meta-reviewing and Sub-PC's) and the series of talks and contributions combined in a section on how to write a good ECIR paper.

Report Feedback

If you have any feedback or comments you would like to raise about ECIR, then please feel free to get in touch with the BCS-IRSG ECIR co-ordinator (Iadh Ounis), chair (Leif Azzopardi) or secretary (Andy MacFarlane).

2. Questionnaire Responses and Feedback

A short survey was conducted before the workshop about ECIR, specifically concerning the structure, organization, and reviewing process of the conference. The survey was sent to 25 information retrieval researchers associated with ECIR (such as those who have ran the conference, served on the programme committee, regular attendees of the conference), of which we received 18 responses. The results are summarized below; issues and questions arising from the survey were discussed during the workshop. The questionnaire is included in appendix A.

Questions and Responses:

Q1 ECIR is becoming increasingly competitive. What should be our response?

Most participants are willing to accept the use of parallel sessions if used to include more highly ranked papers. There was also a preference for parallel sessions over shorter presentations; however it was generally felt that there should be at least one day without parallel sessions. A few participants thought the format was okay as it stands and that no change was needed.

Q2a Should we increase the number of papers?

Most responses were in favor of increasing the number of papers accepted and presented at ECIR. However, this was on some conditions; (i) That more papers are accepted if the quality of these papers is sufficiently high, (ii) if the ratio of acceptance falls below 1:4 – 1:5, and (3) that papers are selected on the basis of overall rank/quality and not over coverage of topics.

Q2b What should be the average number of papers assigned to a referee?

Most participants preferred about three to four papers to review. A quarter of participant would be willing to review five to six papers, and one participant was happy with more than six papers. It was commented that when more papers are assigned, then it is harder to do a good job. Alternatively, if more papers are assigned to a referee, then the Chair is provided with a broader view of the quality of papers. Also, if allocated more papers, then reviewers would need more time.

After discussion during the workshop, it was felt that the length of time given to review (about 4 weeks) was sufficient. However, the timing of the reviewing was problematic; and that moving the submission and reviewing dates forward by about a month would be a better solution. This suggestion could not be implemented for ECIR 2007, due to publication of the dates and the short notice, but will be considered for subsequent ECIRs.

Q3 Should we continue using a double-blind reviewing process?

The vast majority of participants believed that Double Bind reviewing should be kept.

Q4 Should referees be able to self select papers?

Most participants were happy for some form of self-selection. The type of selection preferred was mostly based on the abstract; and if they had some ability to express an interest in a paper. It was felt that this would aid the Chair when deciding the allocation of papers. Some comments noted that reviewers should not expect to receive all the papers they select and realize that they may have to review one or two papers which they did not bid/select. Ultimately, the chair is responsible for the allocation of papers. However, reviewers are usually in a better position to decide if the paper is relevant to their expertise.

Consequently, if a reviewer passed on a paper to be reviewed by a third party, the chair should be informed.

Q5 Should we support reviewers?

The general response was that reviewing support should be provided. Most participants felt that more detailed guidelines would be sufficient. Comments from participants included that the taxonomy of IR papers would be useful, but that guidelines may be difficult to formulate. Also, that extra time could be given to reviewers.

Q6a Should we introduce meta-reviewing?

Half the participants were in favor, whilst half were not in favor of introducing meta-reviewing. Some of the comments suggested that meta-reviewing may introduce bias (given the pool of meta-reviewers to choose from and the necessary rotations required). There was a perception that meta-reviewing introduces a class system of first and second class reviewers. Also, it was noted that the introduction of meta-reviewing would require more time due to the required additional checks.

Q6b What should be the role of the meta-reviewer?

Most participants believed that the role of a meta-reviewer should be to summarize the reviews, ensure agreement and fairness, and maintain ethos of ECIR. However, few thought their job should be to rate/rank the paper! Also, a few participants thought that the meta-reviewers should bring the chair's attention to any discrepancies and issues surfacing from the reviews and try to resolve them.

Q7a Should we introduce sub-PCs which rank/grade a subset of the papers?

Most participants did not believe a Sub-Pc should be employed. However, the comments indicated that participants were not clear of what a "sub-pc" is or entails, and thought that it would be too complicated; and not practical given the volume of papers submitted to ECIR. Other concerns were if a paper is allocated to the wrong sub-pc then it may be disadvantaged. Also, there was perceived problems with calibration between sub-pcs.

Note: during the workshop Gianni addressed many of these pre-conceptions in his presentation on "Sub-PC reviewing" (see later section).

Q7b Role) What should be the role of the sub-PC?

The role of the sub-pc would be to ensure consistency and recommend a subset of the papers.

Q8 Should we obtain feedback from authors on reviews?

Most participants thought that some form of feedback (beyond the reviews) would be useful. 25% for simple rating of review; 25% for comments from authors to PC; 25% for a discussion; whilst 25% were against providing any such mechanism on the grounds that it was not practical and creates more work. It was thought that it would be interesting to try, as it could improve quality, it could be used to reward good reviewers, or could be helpful for selecting the PC for subsequent conferences; however it could lead to frustration because reviewers were unlikely to change their mind; reviewers might feel under pressure and it would also increase the workload of the chair and reviewers (depending on the type of feedback).

Q9a When reviewing, should we make a distinction between student/first timers and standard papers?

About 50% of participants thought that there should be a distinction made when reviewing student papers, but also 50% thought that all papers should be treated equally. Comments for treating papers equally suggested that guidelines for those submitting to ECIR would be sufficient.

Q9b Should we introduce mentoring?

Again mixed results, with half of the participants favoring the idea of providing mentoring, while the other half against. It was commented that it would significantly increase the workload for a few of people (the mentors), the reviews should provide adequate feedback, and in practice it doesn't work well and may introduce bias. However, it may attract more senior IR Researchers and it would be good for EU students.

*Q10 Should there be a preference toward certain types of papers?**Q11 Would you prefer to see papers that are .. [Rank]?*

Q10 and Q11 received mixed results and only attracted responses from a small proportion of participants. We exclude reporting these results. Comments suggested that all types of papers were of equal merit and so long as the papers were of high quality they should be accepted.

Q12 What scale should we employ when grading papers?

Whilst some participants preferred an Odd scale, the majority preferred using an Even scale. The Even scale was preferred because it would force the reviewer to make a decision of accept/reject. Of the different scales the Larger Scale (10) was slightly preferred over 6 or 4 point, as it would give a finer scoring and avoids ties (Also, it can be compressed to five point scale).

What works well at ECIR:

- High student representation, student travel grants, attracts young people
- Venue for discussion (outwit US mainstream),
- Social side, informal relaxed atmosphere
- High standard is maintained
- Small event, growing in size but that is good
- Varied programme

What does not work well at ECIR:

- More transparency and trust in PC/Chairs; review process needs improvement
- The quality of some papers was very low, bad papers are getting in (especially ones that are methodologically flawed)
- Too much like SIGIR in terms of competitiveness
- Limited time to review papers
- ECIR is still finding its identity (friendly vs. formal)
- There seems to be a lack of senior IR researchers attending.

3. Experience from ECIR 2005 PC

The statistics for ECIR 2005 are as follows:

- 24 full papers. 34 accepted (27%)
- 41 posters. 17 accepted (41%)
- 22 student papers

Students were well represented in the conference and there was a good distribution across all areas of IR. Based on the experience of the PC for ECIR 2005, a number of suggestions were made: Extend the reviewing periods; Filter poor-quality papers, perhaps using a subPC to redirect these papers; Have parallel sessions, but not too many; PC chair must be strict with paper submission deadlines, and also with formatting instructions, maximum number of pages per-paper etc.

With respect to the reviewing process itself, the chair could try and combine, where possible, senior PC members with younger researchers, e.g. each paper should be assigned to (at least) one senior reviewer. The chair should ensure that reviewer with different expertise should be combined properly so that the committee has reviews from different perspectives e.g. a paper on logic for Question/Answering should not be reviewed by logicians only. An open issue is how to deal with superficial reviews. Should we give more weight to more extensive reviews? A special discussion on this may be reserved as an agenda item in the PC meeting. The size and composition of the PC should be carefully planned. Last minute decisions should be avoided e.g. bringing in more PC members after being swamped with submissions.

4. Meta-Reviewing¹

There are a number of problems motivating meta-reviewing. Particularly, a very large number of paper submissions require a large number of reviewers to process the papers. As program chairs cannot know the specific expertise of all reviewers, allocation of papers to the relevant expert can be problematic. A large number of reviewers can create problems in that if everyone attends, the PC meeting can be too large and unwieldy, or if few attend there can be an imbalance between subject areas. Quality control can also be an issue if there are too many reviews for the chair or chairs to read, and managing discussions about papers from reviewers can be time consuming. Automatic support is possible to allocate papers to reviewers but manual input is often still needed and there is not automatic support to handle poor or inappropriate reviews.

A solution to these perceived problems is meta-reviewing. This mechanism is popular with ACM conferences and has been used at SIGIR since 1999. An intermediate layer between the PC chairs and PC is created similar to an editorial board for journals. There are a number of different models, but generally a group of ‘Senior’ reviews are delegated to specific tasks.

In SIGIR, meta-reviewers are known as area coordinators and currently complete the following tasks: overview the reviewing of a group of papers corresponding to one of the conference themes; initiate and negotiate discussions between reviews for papers with conflicting opinions; monitor the appropriateness of reviews; provide a summary of discussion and provide independent review of papers; attend the PC meeting. The overall management of papers works in the following way: SIGIR chairs choose PC members, allocate reviewers and papers to area coordinators and chase late reviewers. Different models for meta-reviewing are possible, e.g. lighter/heavier models with more or less involvement from the PC chairs; meta-reviewers can be responsible for a group of reviewers rather than a group of papers.

The perceived advantages of meta-reviewing are as follows. Firstly, a continuity of reviewing protocols can be achieved - depending on good communication between past and current meta-reviewers – so that area coordinators can act as ‘torch-bearers’ for the reviewing process. Poor reviews can be spotted and, if appropriate, revised before the reviews are sent to authors. Expert opinions on all papers is more likely to be achieved; each meta reviewer has a much better idea of what constitutes a good paper in a particular area, rather than what makes a good SIGIR or ECIR paper. The PC meeting is shorter but, hopefully, more accurate. Each meta-reviewer has a specific named responsibility, making task allocation clear and this, in turn, makes it easier to deal with complaints about the reviewing of individual papers.

However meta-reviewing is not without its difficult decisions. An important decision is who selects the meta-reviewers and on what basis – is being an expert in a particular area sufficient or should meta-reviewers be experts in the conference as well. Different conferences have different flavours and audiences might expect particular styles of contributions – a reviewer with experience of a particular conference can be aware of how best to advice authors on this aspect. The personal opinions, viewpoints and beliefs of meta-reviewers can be important in the final decision on a given paper and the relative balance between meta-reviewers over tier reviewers is an important factor. How often the position of meta-reviewer is rotated is also important because meta-reviewers need to be given sufficient responsibility for it to be a sensible task, but sufficiently qualified people tend to be very busy.

¹ This is a personal reflection on the general process of meta-reviewing and does not reflect the views of any other organization or organizing committee.

Meta-reviewing is a sensible approach if the number of submission to a conference is high and a PC chair might struggle with all the tasks they need to complete. Delegating responsibilities to a small group of experienced people can help with the overall quality control of the reviewing process, that is, if the conference can persuade a sufficient number of experienced people to act in this role. However if the number of submissions is low then meta-reviewing is perhaps an additional duty and a big waste of everybody's time. PC chairs may consider using meta-reviewing based on the roles of the meta-reviewers e.g. lighter duties that improve reviewing practice but do not affect the final PC decisions.

5. Reviewing with Sub-Programme Committees

The major objective in reviewing papers is to rank the best papers in a cooperative setting. We use a number of ideal hypotheses or desiderata that support cooperation among reviewers:

- *Peer-to-peer*: the author's knowledge is as good as the reviewer's knowledge.
- *All reviewers are equal*: the role of the PC chair is to select good reviewers, as they were potential authors of papers for the Conference.
- *Review rating*: it should be independent of the expected rank of the paper and acceptance rate of the Conference.

There are a number of problems with ensuring cooperation; in practical situations the acceptance of a paper may cause a conflict, and thus a rejection of another one. The necessity of introducing a ranking for the definition of the final programme is affected by two major problems. Reviewing ratings can be inconsistent between areas and sets of reviewers, and normalising these ratings can be difficult. In addition, different schools of thoughts or different subjects may work in a competitive way.

With regard to the best or 'perfect paper', what is desirable? A number of opinions can be expressed about a paper. When we enjoy reading a paper we usually express some preliminary thoughts: "I wish I had written this paper!", "I now understand the issue discussed much better", "The method suggested in the paper works", "How does the method suggested connect with my approach; can I use it?", "This method can be applied to a different problem.", "This paper made me think.", "I have a different perspective of this problem.", etc.

What would be the perfect model for a PC? All members of the committee would be cooperative. All members of the committee would read all papers, and they would all attend the PC meeting to finalize and discuss the conference programme. In practice, this does not happen. There are two phases (i) reviewing, which is not cooperative, and (ii) finalizing, which is cooperative among a subset of the PC who attend the final PC meeting. There can be a problem in the finalization stage in recovering the missing knowledge about the papers on the borderline, because in general the highest ranked papers are not examined, and thus the PC attendees have not a clear and general view of the final programme of the conference.

In order to resolve these problems we can consider the use of a Sub-PC e.g. the employment of the perfect PC model on a subject of the papers submitted to the conference. In this model, reviewers of papers in the same area probably need to have a general vision of the submitted papers. The procedure is to assign a cluster of papers to a cluster of reviews, instead of assigning a single paper to a single reviewer at random, provided that the paper content matches the reviewer profile. In the Sub-PC model, a subset of the PC reads a subset of the papers, but all reviewers rank all the papers. This is equivalent to having a meta-reviewer, but Sub-PC is different to the concept of meta-reviewing because it satisfies the peer-to-peer axiom. Also the necessity of rotation of meta-reviewers is not needed.

Unlike meta-reviewing, a Sub-PC does not require changes in organization from previous instances of the conference and reviewers may not be even aware of the existence of the Sub-PC in the reviewing process. This is useful if time is tight. On the other side, there is more effort for PC chairs to provide a good covering with a good Sub-PC, taking into account that members of the Sub-PC should be proportional to the number of papers per-area. A fair and normalized ranking on the set of papers is more likely with sub-PCs. Biased reviews can be detected more easily and the biased clusters of reviews can be detected by statistical means. To ensure a sound reviewing process, each cluster may have an experienced reviewer. Reviewing load is an important constraint on this methodology; the number of members in a Sub PC (X) and the size of the set (Y) of papers assigned to it cannot be large. For example, in ECIR 2007, we have singled out 11 macro-areas, which means an average of 2-4 sub-PC's for each macro-area, depending on the number of papers submitted for macro-area.

Consider the following example. We have 250 papers submitted; with three reviews each requires a total of 750 reviews. Assume we have 120 PC members in total. Given this $\#X = 120/3=40$ PC clusters $\sim 250/6\sim 40$ paper clusters = $\#Y$. One paper per-cluster will be accepted for presentation. The sub-PC's model thus introduces two types of borderlines (cluster and ranking). If the distribution of papers is random, the first ranked paper of a cluster should belong to the set of the highest ranked papers. If there are too many accepted papers from the same cluster, the PC meeting will discuss these clusters. Any contiguous clusters could be investigated further. In this way, the PC meeting will have a much better and more organized vision of the final programme.

6. How to write a good ECIR paper

Introduction

The purpose of this section is to provide a set of guidelines for authors wishing to submit to ECIR. The goal is to provide an overview of the different types of papers that are acceptable and what is required given that type. We provide a guide for the following areas: theoretical, experimental/comparisons, People in IR, applications, conceptual, and evaluation and performance measures. The guide is also intended to aid reviewers of ECIR papers, as well to provide an indication of the expected level of work and the types of acceptable papers.

Properties of a good IR paper

A good IR paper (or any scientific paper) for that matter should engage the reader and succinctly convey an original idea. The contribution to knowledge should be expressed as clearly as possible to facilitate understanding and provide the reader with a better insight. For instance, the paper should aim to provoke such reactions from the reader along the lines of, “I now better understand this issue.”, “It works!”). Or enables the reader to draw connections between other bodies of research (“I’ll see how this fits with my XYZ approach”, “It can also be applied to ABC!”). Or invokes the reader’s imagination or changes their perspective (“This paper has got me thinking!”, “I haven’t seen this from this point of view before.”). A great IR paper will obtain a response from the reader like “Why didn’t I write this paper!”. Of course, not all papers invoke such responses, for various reasons. Here, we focus on some of the main genres of papers that are encouraged to be submitted to ECIR, and describe some of the qualities that they should possess.

Please note that these are guidelines and authors and reviewers should use them in the way intended; as guidelines to help the process of writing and reviewing. Consequently, the author and referee should use their experience, common sense and discretion when interpreting and using these guidelines.

Theoretical Papers

What is a theoretical paper?

A theoretical paper proposes a theory for Information Retrieval (or some phenomena within the domain). A theoretical paper should present a supposition or system of ideas intended to explain some phenomena within IR. It should be based on general principles independent of the phenomena to be explained (i.e. Darwin’s Theory of Evolution). It could provide a set of principles on which the practise of an activity is based (i.e. a theory of information seeking behaviour). Or it could present an idea used to account for a situation or justify a course of action. Consequently, this does not necessarily imply that the theory is grounded in mathematics or some other formalism, which is common misconception about theoretical papers in IR. Instead, a theoretical paper may also be discursive in nature, providing arguments and reasoning through the discourse.

What makes a good theoretical paper?

First, the paper must go beyond the existing theory already present in the literature – and thus fulfil the originality criteria. In order to convince the reader that this is the case, there should be links to older theory to provide the context of the paper. The relationship between the old and the new should be related and explained.

Second, it is important for a theoretical paper in IR to provide the necessary contextualisation of the theory within IR. That is, what is the relevance of this theory to IR? Consequently, the generic application of a machine learning approach, for example, is not

relevant. The burden is on the writer to exemplify the link between the theory and the practise, given the domain.

Third, the clarity of the presentation is very important, because the emphasis of the paper is to present an account for a phenomenon. Consequently, the arguments presented need to be clear and justified. One way of ensuring clarity is to provide illustrative and practical examples to aid the reader's understanding.

Forth, a theoretical paper aims to link theory with practise; once a theory is presented, the inevitable question arises; does it work in practise? However, "proof" that a theory holds is not a necessary requirement for a theoretical paper to be acceptable, as it is not always possible for a theory to be put forward and for it to be tested to the nth degree. There are various reasons for this; the work is in its early stages; the machinery doesn't exist for it to be tested; etc.

In such cases when experimental work can not be provided to ensure that the paper is acceptable, there are other criteria that the paper should meet. A discussion should be included about the testability of the theory presented, comments on whether it can be falsified, its tractability, how the theory could be tested in practise, its relationship with experimentation, and whether it is possible to implement or not. Addressing such issues is paramount to papers, which present novel/new theory.

However, there are cases when the theory presented is an extension to the existing theory. In this case, where the theory has been tested previously, it is necessary to provide some experimental work in order to show that this extension is actually significant, useful, successful etc. Re-stated, delta theory papers should provide some empirical testing. On the point of significance, a theoretical paper should also discuss what would constitute a significant result and how to quantify this.

Experimental/Comparisons Paper

What is an Experimental paper?

An experimental paper compares one or more competing theories/techniques within Information Retrieval. An experimental paper should contain the context of study, a clear statement of the problem addressed, and present clear research hypotheses.

What makes a good experimental paper?

The paper should make an original contribution to IR and state clearly what exactly is new with respect to previous work. Consequently, a good set of references should be included to link prior work; and should include those approaches, which can be used as a baseline. An experimental paper should use publicly available and (preferably) standard test collections. For instance, the generally accepted collections for empirical evaluation are those provided by TREC, but also include collections from other common evaluation forums such as CLEF, INEX, etc. The use of older collections such as Medline, Cranfield, CACM, NPL, etc are now considered too small to be acceptable. Experimental papers that use parts of test collections or subset of topics are generally considered unacceptable, unless accompanied by a reasonable justification (see below).

The use of a non-standard test collection can be acceptable if it is publicly available, and representative or diverse enough to warrant reasonable conclusions. Note that the requirement of test collections being publicly available is to ensure that the experiments performed within the paper can be reproduced. Consequently, a good paper will ensure that the data is available to enable replication, verification, and/or reproducibility of the work. If the data used in experiments is not publicly available then the findings, reports and conclusions drawn should not be based on such data. For instance, such experimental analysis may complement the findings shown on publicly available data collections. Or, it could be used to illustrate or illuminate the findings but are not pinnacle to the contribution.

An experimental paper should justify the data collection(s) and analysis methods used. Depending on the retrieval task, the paper should use an appropriate test collection (generally the most recent one), all its associated topics and assessments coupled with a

suitable analysis method. In particular, the use of a non-standard test collection should be justified and ensure that there is access to the collection (either, through the author, on their website, etc). A good experimental paper should use more than one test collection (if available) to provide more evidence for the hypotheses presented and show how generalisable the techniques examined are.

An experimental paper should use appropriate statistical or qualitative methods and report appropriate and standard evaluation measures. However, simply performing and reporting significance test and so forth is not sufficient without further explanation of that significance (see below).

Importantly, an experimental paper should use appropriate state-of-the-art baseline(s) to convince the reader that the proposed technique is superior or not. The characteristics of a good baseline include that it provides strong retrieval performance, is well-established and robust across different collections. Other baselines may include the best performing runs at TREC, for instance. Though, they should not be used as a sole indicator (as such runs have not been optimized/tuned given the collection). If tuning or optimizations of the techniques or models are required then this luxury should also be afforded to the competing baseline models. If the experimental paper proposes a technique, which is computationally expensive, this inefficiency should be acknowledged and discussed.

Finally, an experimental paper should indicate the significance of the results and conclusions made with respect to the practice and/or theory of IR. By reflecting on the methods, datasets and results (or combination there of) the significance of the work should be provided to explicitly and precisely describe how the work is better/useful/interesting/etc with respect to the current start-of-the-art. An excellent paper will provide insights into why it is succeeding or failing and also discuss how generalisable the results are. A key point, here, is that a paper in this genre is still acceptable even if the model does not perform as expected, so long as this insight is provided. Conversely, a paper presenting outstanding results should not be simply accepted on the basis of the superior results, if there is no rationale and understanding to how these results were achieved.

“People in IR” - User studies/Interfaces papers

What is a “People in IR” paper?

“People in IR” type papers cover a variety of research within Information Retrieval; the distinguishing feature of these papers is that they involve humans as a major component in the system, experiment or investigative study being described.

Broadly, there are two main types of “People in IR” papers: (1) those based on laboratory IR investigations – these are similar to experimental papers but with the involvement of humans, and (2) those investigating information seeking and behaviour. The first type, generally referred to as Interactive IR, includes evaluation of novel interfaces, and interaction work, including user modelling and predictive or adaptive technologies (with people involved in the evaluation or data collection). The second type, dealing with what is commonly called Information Seeking or Information Behaviour, is more concerned with the information needs and search behaviour of individuals or distinct groups of people. Papers in the Information Seeking area may be more discourse-based than is normally seen at ECIR. Both types of research are, however, welcomed at ECIR.

There are core approaches and techniques to facilitate research in the area of interactive IR and Information Seeking. However, there are a few fixed methodologies. Due to the variety of research within this genre, a “people in IR” paper usually describes a novel methodology specifically created for an individual investigation. This methodology – a coherent set of decisions and investigative components – and the reasons behind the methodology will require explanation within the paper. This is in contrast to other branches of IR where we have standard methodologies, metrics and tools such as test collections that are commonly accepted within the discipline and require fewer introductions.

What makes a good “People in IR” paper?

Any good “People in IR” paper should provide a coherent narrative to describe the research questions motivating the research, the methodology to investigate these questions – including the design decisions behind any novel system or interface development – the relevant results obtained and the implications for future research.

A laboratory based IR paper is, in many respects, similar to experimental papers and the same general guidelines apply. The introduction of people within the studies however introduces some further issues. In particular, the paper should describe the people involved and why their particular characteristics, such as search experience, might influence the results obtained. Similarly, the paper needs to describe the components of the study, such as the source of search tasks, any baseline systems, and instructions given to participants in the study and how these aspects relate to the research questions and results obtained. It is very important that a paper shows these connections to convince the reader that the experimental set-up is not unrealistic or biased.

An Information Seeking paper should also explain the methodologies chosen to investigate the research goals of the paper and, where appropriate, discuss alternate methodologies that could also have been followed. A good Information Seeking paper will investigate search phenomena in depth rather than just reporting or describing the study and basic results. Rather, a good paper will seek to investigate the reasons for the results and will present and analysis of the implication of the findings for IR research.

For both Interactive IR and Information Seeking papers, it is important that the evaluation and methodology should be appropriate given the research hypotheses and objectives. Common criticisms of such papers are that there are not enough participants, user groups, tasks, or baselines. However, such criticisms should only be made with respect to the methodology and research questions presented. For example, an evaluation of an interface that is intended to be used by a wide group of searchers for all search tasks will require more participants and more search tasks before we can accept that the results obtained are meaningful. A more qualitative investigation on, for example, children’s uptake of new search technologies, may require fewer participants because each participant will be analysed in more depth. Consequently it is important to justify such design choices and also to explicitly acknowledge any limitations of the study and how such limitations might affect the outcomes of the study.

In both types of study, many results will be obtained, all of which cannot be presented in one paper. Therefore it is necessary to select results rather than overwhelm the reader with as many as possible. To avoid appearing to be overly selective (i.e. only presenting results that are positive with respect to some existing hypotheses) it is better to concentrate on a smaller number of related results and research questions and investigating them in more depth. Unexpected or surprising results are worth including as is qualitative information from any participants in the study. Qualitative information helps contextualise the quantitative results and helps the reader understand the experience of the human participants.

What makes a poor “People in IR” paper?

A weaker paper in both Interactive IR and Information Seeking can suffer from a number of problems that make it hard for the reader to appreciate and follow the research being presented. A common flaw, and one that will kill most papers, is poor exposition of the research itself – not explaining why the research is being carried out, how the study was constructed, which individual results were selected for presentation, and how the research links with other research in the area. A particular weakness for Information Seeking papers, when submitted to a conference such as ECIR, is not to discuss the implications of the work for the general field of Information Retrieval – how might the research being presented change the way IR systems or interfaces are designed or evaluated?

Methodologies for evaluation or investigation of behaviour are usually complex and we cannot explain every detailed aspect of the methodology within the limits of a conference

paper. However, it is important to provide sufficient detail so that the reader can follow the study being described but also the thinking behind the design of the study.

A second and common problem, particularly for Interactive IR papers, is to simply present the quantitative results rather than describing them and their importance. In a test collection evaluation it *may* be acceptable simply to show statistically significant results between standard evaluation metrics in order to convince the reader of the benefit of one algorithm over another. In an Interactive IR paper, this is usually not enough; rather the results have to be explained to the reader. If searchers run more queries on a novel interface than on a baseline, for example, the author should explain why this behaviour might have occurred with reference to the design principles of the interfaces and preferably to the other results obtained within the study to present a full understanding of the importance of the result.

The most common flaw with Information Seeking papers is simply to present the author's experience of running the study with no attempts to validate their findings (e.g. by using a mixture of elicitation methods), no attempt to discuss the implications of the research and no attempt to relate the study to existing work in the literature.

Applications and System Prototype papers

What is an applications paper?

There are four main types of application papers:

- (1) Positioning papers;
- (2) Technical papers;
- (3) Demo papers; and
- (4) Test and evaluation papers.

A positioning paper details the motivation and background for an application. A technical paper details the description of the architecture, individual components, algorithms, integration of components, etc. A demo paper describes the system in the paper and is usually coupled with a demonstration of the application. A test and evaluation paper reports empirical results from the testing of an application.

An applications paper, therefore, may report different stages or phases of a research project and an application prototype and its development; such phases include requirements and design analysis, Prototyping and implementation, testing and evaluation, and dissemination. An applications paper may iterate across the above phases to produce papers of type 1 to 4 respectively. Obviously, the expected contribution and impact of the paper will/should grow as later phases are reported.

What makes a good Applications papers?

An applications paper should include an explicit system description and take the reader through a user experience or scenario, if applicable, providing examples such as a walk through of the system and the iterations.

It is important that an applications paper contextualises where in the timeline, the application is at in the series of phases. Further, this contextualisation should also include how the work relates to the system as a whole. To facilitate this, it is important that related work is cited in the relevant disciplines. As a consequence, the reader can fully appreciate and contextualise the work, and the contribution's references to any prior work should be included.

A positioning paper should present a thorough and comprehensive background along with a detailed motivation for the application. The uniqueness of the solution/application should be explained and the justification for how this position was arrived should be provided.

A demo paper should provide a description of the system and how the system would be experienced by the user. Since it is a demo, there is an expectation that presentation will

contain a demonstration of the application/prototype system; i.e. the system description is of the demo, rather than what is to be built, and the system should be working such that reasonable feedback can be given by audience. The system description should include the science and motivation behind the application to justify why the application is novel and warrants demonstration. Along with the system description should be a technical specification stating the configuration, hardware and other requirements that the application requires. This should be done separately from the system description to avoid oscillating between motivation/science and the technical aspects.

In a Test and evaluate paper, a clear distinction should be made between the testing of the system (through running experiments, etc) and the evaluation and analysis of the experiments. During the test process, the inclusion of a functionality check should be included to detail what is operational in the application and what is not, and to specify any other limitations relevant to the experimentation. As with user studies papers, a good applications paper of this kind, will describe the experimental conditions under which the user testing took place; for instance pointing out whether real users were involved or whether it was pilot tested on colleagues.

What makes a poor applications paper?

An applications paper that presents an idea (new or old) but no evidence or explanation of its perceived need or uniqueness is not acceptable. A paper, which is a technical push of some technology, is not acceptable, unless a clear requirement for the specific technical improvement is shown and justified, or if a novel vision is presented where the technical push would be appropriate. Application papers should provide sufficient evidence to motivate or justify their arguments. Consequently papers which rely on hearsay, word of mouth, or practise which has become a de-facto standard, do not provide adequate justification.

Conceptual papers

What is a conceptual paper?

A conceptual paper presents concepts dealing with or relating to IR, where a new perspective is obtained or formulated by the combination of a group or class of objects. For instance, the identification of trends or patterns, which occur in IR, where the contribution to knowledge is the definition of the concepts and their relationships to the IR process. An example of a conceptual paper is one that defines a new way to characterise the notion of relevance in IR, and shows how this new way links to other, and also what it brings to what was there before.

In general, a conceptual paper is likely to be discursive. This is to our point of view what makes a conceptual paper different to a formal/theoretical paper, although the boundaries here can be very fuzzy. Also a conceptual paper requires a very deep understanding of the problem or issue being addressed, investigated, studied etc and as such we would say that such papers are difficult to write.

What makes a good conceptual paper?²

A good conceptual paper should state and formally define all the concepts introduced and their relationships/interactions. This can be thought of as the conceptual development.

The conceptual development should strive to maintain consistency in the level of abstraction and the unit of analysis. Also, when using multidimensional concepts (for example in INEX, relevance is defined as a two-dimension multi-graded concept), all the relevant dimensions should be clearly explained. The underlying assumptions of the concepts should also be

² This section has been based on the AMS Review guidelines for conceptual papers, see [AMSR].

explicitly stated (e.g., conceptual paradigm), along with the boundaries or limitations of the conceptual development. Consequently, a good conceptual paper will also explain for whom and under what conditions the proposed conceptual development is appropriate or inappropriate.

Not only must conceptual paper justify the conceptual development, but it should also show why this is a viable way to do so, as opposed to other ways. Hence, a good conceptual paper establishes a clear link between the proposed conceptual developments and previous research (including alternative perspectives) and explains how the conceptual development improves upon existing conceptualizations. This is very important, and requires a strong knowledge of the problem/area by the author.

An important part of the contribution made by a conceptual paper is how the proposed conceptual development changes our understanding of the theory and practice in IR. In other words, a conceptual paper should spend considerable effort establishing the implications of the developed concepts, which are being proposed.

Similar to theoretical papers, a conceptual paper should make certain that the conceptual development can ultimately be tested and measured empirically.

A bad conceptual paper is one that contains lots of (eventually) interesting concepts, whether novel or not, but with very little connection to the IR problems being tackled. Defining concepts and their relationships, however elegant, is not enough. We must make sure that we relate the proposed conceptual development to the problems being addressed, and how they relate to state-of-the-art work.

Evaluation and Measurement papers

[To be added]

What is an Evaluation paper?

What makes a good Evaluation paper?

Taxonomy of research papers as applied to IR

According to Parberry [parberry94], research papers can be broken in six main categories: Breakthrough, ground-breaking, progress, re-prise, tinkering, debugging, and survey. Here are the categories with respect to Information Retrieval.

Breakthrough

This is typically an open problem that has persisted for some time, which has received a considerable amount of research effort dedicated to solving that problem. (e.g. dealing and handling context, evaluation, introduction of pioneering models, i.e. language modeling [ponte98lm], BIM [robertson76bim], etc.)

Ground-Breaking

A paper would fall into this category if it opens up a field within IR that is not well explored or understood, and provides a firm foundation for further research. (Examples, vector-space model [salton68vsm], inference networks [turtle90in], non-classical logics [vanrijsbergen86nc], etc.)

Progress

This type of paper would solve or address a new or recent open problem, typically limited to

that particular context. (Apparently, most papers are of this variety.)

Reprise

A paper of this variety would provide superior evidence (possibly contradictory evidence) of a previous result. It is important to ensure that there is both elegance and insight obtained by the paper. This could entail the use of a better experimental design to test the hypothesis, which results in more conclusive evidence, an analysis that is more thorough than the previous, makes more illuminating connections than the past work. (Experiments contesting the cluster hypothesis, would tend to be of this type, for instance)

Tinkering

Such papers are only really of merit if the extension of known results is provided through a more careful and detailed analysis, but in a non-obvious way. (As opposed to incremental research, tweaking the algorithm for performance increases etc)

Debugging

Such papers would elucidate and then repair a previously undiscovered flaw in previously published work.

Survey

A paper in this category would unify the particular area of a specialized subject within IR with consistent notations, terminology, etc, often piecing together results from many disparate sources. For instance, Hawkings' paper on Enterprise Search [hawking04es] or Callan's paper on distributed IR [callan00dir].

Discussion of issues Raised during the Workshop

There were a number of issues discussed during the workshop, which needed to be raised for the next ECIR, to be held in Rome in April 2007. A number of key decisions were made as follows:

- **Meta-reviewing vs. Sub-PC:** A Sub-PC would be utilized for ECIR 2007. It was felt that the introduction of meta-reviewing would not be appropriate because it would increase the reviewing turn around time, requires a large subset of meta-reviewers to drawn upon, and the number of submission is probably not high enough to warrant the overhead. On the other hand, Sub-PC was preferred as it has the potential to address the concerns of quality, without disturbing the current processes. That is, the reviewing remains the same, except the allocation of papers by the chair is organized into Sub-PCs.
- **Length of Conference.** The decision to have a longer conference, utilize parallel sessions was deferred until another time. Given constraints in venue and the number of participants expected, ECIR 2007 will be held on 4 days. One reason put forward for having a longer conference was that it enables the participants to see more of the conference (as parallel sessions are minimized) and that makes the sessions more relaxed (as participants are rushing in and out). However, the discussion revealed that, generally, extending the conference from 2 and half days to three full days would be acceptable. It was felt that four days would be too long among those present at the workshop.
- **Reviewing Guidelines:** The presentations about how to write a good IR paper along with discussion formed the agreed set of draft guidelines (See above).
- **Delegation of reviews:** One issue raised in the questionnaires was the delegation of reviews by referees. For instance, there was concern that papers may be farmed out to PhD students without them getting recognition or training. It was discussed and agreed that if a paper is delegated then the referee to whom the paper was initially assigned needs to contact the PC Chair as it is the responsibility of the PC Chair to delegate the reviews, not the referees. Any additional reviewers should be named. The draft guidelines set out should also provide some guidance for reviewers of ECIR.

Other points discussed at the workshop:

- After discussion of double blind reviewing it was agreed to keep the anonymity as is.
- The acceptance rate was discussed (i.e. whether to accept more papers), it was agreed that around 35-40 accepted papers is still appropriate. However, this can only be a guideline as quality should take precedence in the paper's acceptance.
- The introduction of short and long papers was briefly discussed, but without any firm resolution. One point that did emerge was that if short papers or posters are accepted, then it might help to also provide a guide on how to review such pieces of research. (Note that posters are currently accepted, however, the same guidelines for papers are used for posters).
- The problem faced by reviewers with respect to weak or off topic papers was discussed. To what extent should these papers be reviewed? It was agreed that both of these issues would be dealt with at the discretion of the PC. Papers which are either

submitted in the incorrect format or are over lengthy will not be reviewed, and will be automatically rejected. Any PC chair or member who would like clarification on reasons for rejecting a paper should refer to [Parberry95].

- When accepting a paper, it was strongly recommended that reviewers include comments on how the paper could be improved for the conference proceedings.

Other decisions: It was felt that there is no strong reason or need to run a doctoral consortium along with ECIR as the conference is aimed at providing a forum for young researchers. Nor was it felt that workshops should be added to the conference at this time, though it may be worth considering in the future.

References

[Parberry94] Ian Parberry. *A guide for new referees in theoretical computer science*. Information and Computation, 112(1), 1994.

[AMSR] <http://www.amsreview.org/submit.htm>. Academy of Marketing Science Review, Submission Guidelines.

[ponte98lm] Ponte, J. M. and Croft, W. B., *A Language Modeling Approach to Information Retrieval*, In the Proceedings of the Twenty First ACM-SIGIR, 1998.

[robertson76bim] Robertson, S. E. and Sparck-Jones, K., *Relevance weighting of search terms*, Journal of the American Society for Information Science, Vol, 27, 1976

[turtle90in] Turtle, H. and Croft, W. B., *Inference Network for Document Retrieval*, In the Proceedings of the 13th Annual International ACM SIGIR, 1990

[vanrijsbergen86nc], van Rijsbergen, C. J., *A non-classical logic for information retrieval*, The Computer Journal, Vol. 29(6), 1986

[salton68vsm], Salton, G. and Lesk, M. E., *Computer Evaluation of indexing and text processing*, Journal of the ACM, Vol .15(1), 1968

[callan00dir], Callan, J., *Advances in information retrieval, Chapter5: Distributed Information Retrieval*, 2000

[hawking04es] Hawking, D., *Challenges in enterprise search*, In the Proceedings of the fifteenth Australasian database conference, 2004

Appendix A – Survey Questionnaire

ECIR Review Workshop Survey

The purpose of the workshop is two-fold. First, we want to revise the organization of the PC and second we want to set up some guidelines for what makes a good (EC)IR paper (and the types of papers that can be submitted). The main reason for the workshop is the significant increase in the number of ECIR submissions. Developing procedures and guidelines to deal with this increase in quantity, whilst maintaining the ethos of ECIR and the quality of reviews and papers is the main goal.

This survey is intended to help guide the first session of the workshop so that we discuss the most pertinent issues with respect to the organization of ECIR (therefore most questions are related to the first aspect of the workshop). Please keep in mind that not all suggestions and opinions can be implemented due to various constraints (time, cost, etc). Please feel free to add extra comments or opinions on the various aspects of the reviewing process.

- (1) ECIR is becoming increasingly competitive (high submission numbers, low acceptance rate). ECIR's response to this change should be:
- (a) make more use of parallel sessions to include more high ranked papers
 - (b) have shorter presentation times to include more high ranked papers that everyone can hear
 - (c) have a longer conference
 - (d) do nothing and have only the best papers presented to the whole audience
 - (e) other (please specify)

comments:

(2a) The currently accepted number of papers is around 35. Should we increase the number of papers?

- no
- yes
- if yes, why?
- If no, why?

(2b) What should be the average number of papers assigned to a referee?

- 3-4
- 5-6
- 6+

comments:

(3) Should we continue using a double-blind reviewing process?

- yes
- no

comments:

(4) Should referees be able to self select papers?

- Yes, on full paper.
- Yes, on abstract.
- No

if yes, how would those papers not selected be handled?

comments:

(5) Should we support reviewers?

- yes
- no
- if yes, what support should we consider? Reviewing workshops? Guidelines? etc.

Reviewing Organization

Potential ways to address the quantity and quality issue are through the introduction of meta reviewing and/or sub PC's.

(6a) Should we introduce meta-reviewing?

- yes
- yes, under certain conditions...
- no
- no, we should introduce another way to handle the quantity/quality...

comments:

(6b) What should be the role of the meta reviewer? (check *all* those that apply)

- rate paper
- summarize reviewer comments
- ensure referees agree
- ensure the reviews are fair
- ensure the reviews maintain the ethos of ECIR
- other

comments:

(7a) Should we introduce sub-PCs which rank/grade a subset of the papers?

- yes
- yes, under certain conditions...
- no
- no, we should introduce another way to handle the quantity/quality...

comments:

(7b) What should be the role of the sub PC? (check *all* those that apply)

- provide a consistent rating for the set of papers
- recommend a subset of the papers the sub PC reviewed
- other

comments:

(8a) Should we obtain feedback from authors on reviews? (Check *all* those that apply)

- allow discussion
- allow discussion, but limited only to the reviewers comments (i.e. the right to reply)
- comments from the author to the pc only about the reviews
- simple rating reviews (was this review helpful 1-10?)
- no discussion
- other

comments:

(8b) If, we to obtain feedback from authors, when should we obtain author feedback?

- before pc meeting?
- on notification of acceptance/rejection?
- during the reviewing process (i.e. as the reviews come in)

comments:

(9a) When reviewing, should we make a distinction between student/first timers and standard papers?

- student papers, only
- first time authors at ECIR
- no, they should all be treated equally
- if yes, what distinction should be made?

(9b) Should we introduce mentoring?

- yes
- yes, but only for first timers/students
- no
- other, please specify

comments:

Rating Papers

(10) Should there be a preference toward certain types of papers? Rank

- Conceptual
- Theoretical
- Applications and system prototyping
- User-Studies and Interfaces
- Surveys
- Experimental/ System Comparison
- Evaluation and measuring performance

comments:

(11) Would you prefer to see papers that are? Rank [See appendix below for definitions]

- Ground Breaking
- Break Through
- Survey
- Tinkering
- Debugging
- Reprise
- Progress

comments:

(12) What scale should we employ when grading papers?

- 4 point
- 5 point
- 6 point
- 10 point
- other, please specify

comments:

Comments and Open Questions

(13)

- What works well about ECIR now?
- What doesn't work well about ECIR now?
- Are there any other subject areas that would improve relevance to state of art work?
- Should we have a Doctoral Consortium?
- Can you think of ways to simplify review process?"
- Are they any issues you would like to mention, raise?