InexBib - Retrieving XML elements based on external evidence

Alexander H. Krumpholz
David Hawking
Information Retrieval Group
CSIRO ICT Centre
Canberra
Alexander.Krumpholz@csiro.au

Abstract

Creating a scientific bibliography on a given topic is currently a task which requires a great deal of manual effort. We attempt to reduce this effort by developing a tool for automatically generating a bibliography from a collection of articles represented in XML. We evaluate the use of elements around the references as anchortexts to improve search results. We find that users of the tool prefer lists generated using anchortext over those generated from the bibliography entry only and that the preference is statistically significant. We tentatively find no significant preference for results generated using paragraph as opposed to sentence level anchortext, but note that this finding may result from lack of sophistication in resolving text including multiple references.

Keywords Information Retrieval, XML, Element Retrieval, Bibliography

1 Introduction

Over recent years XML has become a standard data exchange and storage format in all application domains. The 'INitiative for the Evaluation of XML Retrieval' (INEX) [5] studies XML retrieval techniques and evaluation methods. Its approach is focused on the retrieval of XML elements specified in the query (Content-and-structure (CAS) queries) or those best matching the search terms (Content-only (CO) queries), but in practice in this application, searchers usually prefer to retrieve whole documents. Indeed researchers have struggled to find motivating examples for element retrieval. We propose the construction of a reference list for a given topic as such a task, using the text around the reference in a publication analogously to anchortext in web retrieval to increase the retrieval quality.

Proceedings of the 11th Australasian Document Computing Symposium, Brisbane, Australia, December 11, 2006.

We present the results of a pilot study comparing the perceived quality of bibliographies generated with and without the use of 'anchortext'.

2 Related work

Previously published work in the areas of bibliometrics and the exploitation of anchortext in Web search are somewhat relevant to the present study as are systems such as CiteSeer[1] and Google Scholar[2].

2.1 Use of anchortext in retrieval

The usage of anchortext (the text actually forming the clickable link on web pages) has long been used to increase the retrieval quality of web search engines [12, 2, 3, 4]. Analogously to anchortext we indexed the text surrounding references as additional text for the bibliography entry. In two experiments we extracted the embedding sentence and paragraph respectively to explore the impact of context sizes on the retrieval quality..

Note that the use of anchortext introduces a voting effect. Bibliography items which are cited multiple times with descriptions matching the query will be ranked more highly than less frequently cited items.

2.2 Bibliometrics and bibliography generation

The field of bibliometrics [6, 14, 7, 10] concerns itself with the graph of citation links between scientific articles and has provided inspiration for link-based ranking methods in Web IR (e.g. [2]) but does not take account the descriptive text in citations.

CiteSeer [1] started to index publications and citations in 1998 and has become a widely used resource. Over time the CiteSeer database has grown to 730,000 documents with over 8 million citations and a new version CiteSeerX has recently been presented [11]. CiteSeer has access to a much larger

[1] http://citeseer.ist.psu.edu/
[2] http://scholar.google.com/

database of citations than we are using, and can be used to retrieve a list of references that match query keywords. However, we are not aware that CiteSeer uses descriptive anchortext in the retrieval process.

No other publication known to the authors investigates anchortext approaches for reference list generation.

3 Method

In this section we characterize the INEX data, and explain how we extracted bibliographic items and matched them to citations in the articles. We then describe the retrieval software we used and how we built the three different indexes used in the study.

3.1 INEX data

The data corpus used by INEX for the last four years is a collection of over 12,000 journal articles from 18 IEEE journals from years between 1995 and 2002. (See tables 1 and 2.)

The articles are stored in XML format as described by Fuhr et al. in [5], allowing researchers to develop and apply XML retrieval techniques to create the result lists defined for the current INEX round.

Listing 1 shows an example reference, Listing 2 an example bibliography entry.

In order to retrieve elements other than those whose bibliography entry matched the search terms, we extracted almost 150,000 references from the bibliographies of all the journal articles in the collection, saved them into separate files and used naive record linkage techniques to identify publications cited by multiple articles.

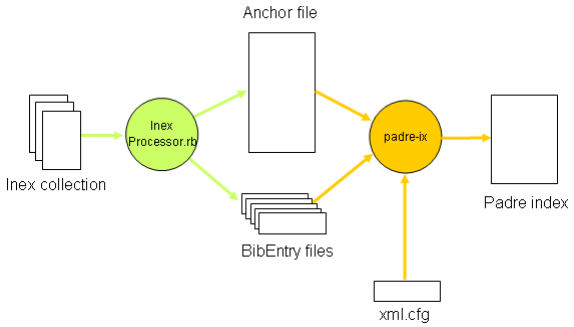

Figure 1: Preprocessing

```
...
A more detailed discussion on
fairness can be found in
[<ref rid="bibL03371" type="bib">1</ref>],
[<ref rid="bibL033710" type="bib">10</ref>].
</p>
...
```

Listing 1: Example reference

```
... <bb id="bibL03371">
  <au>
    <fnm>K.R.</fnm>
    <snm>Apt</snm
  </au>
  <au>
    <fnm>N.</fnm>
    <snm>Francez</snm>
  </au>
  <obi>and</obi>
  <au>
    <fnm>S.</fnm>
    <snm>Katz</snm>
  </au>
  <atl>&amp;ldquo;Apprasing Fairness in
       Languages for Distributed
       Programming,&amp;rdquo;</atl>
  <ti>Distributed Computing,</ti>
  <obi>
    <volno>vol. 2,</volno>
  </obi>
  <pp>pp. 226-241,</pp>
  <pdt>
    <yr>1988.</yr>
  </pdt>
</bb> ...
```

Listing 2: Example bibliography entry

3.2 Record linkage

The INEX collection contains the bibliography entries and the references already extracted into XML elements and linked via artificial keys. However, as the keys are only unique with one journal article, multiple articles refer to the same publication using different keys. Since bibliography entries are often referenced by multiple journal articles, a deduplication problem had to be addressed. In order to prepare for the bibliography entries to be indexed, each entry has been extracted into a separate file.

The document type definition of the journal articles has been defined very vague to allow all possible bibliography entries to be recorded; however, this allows for multiple or missing elements, which made the record linkage a non-trivial task.

We introduced a key for bibliography entries by combining the first author's last name with the title of the publication to identify references of identical publications. However, even though we converted all characters to lowercase, removed special characters and entities like '“' (see Listing 2), some

Period	Journal
1995 – 2001	IEEE Annals of the History of Computing
1995 – 2001	IEEE Computer Graphics and Applications
1995 – 2001	Computer
1995 – 2001	Computing in Science and Engineering
1995 – 2001	IEEE Design & Test of Computers
1995 – 2001	IEEE Intelligent Systems
1997 – 2001	IEEE Internet Computing
1999 – 2001	IT Professional
1995 – 2001	IEEE Micro
1995 – 2001	IEEE MultiMedia
1995 – 2000	IEEE Concurrency
1995 – 2001	IEEE Software
1995 – 2002	IEEE Transactions on Computers
1995 – 2002	IEEE Transactions on Parallel and Distributed Systems
1995 – 2002	IEEE Transactions on Visualization and Computer Graphics
1995 – 2002	IEEE Transactions on Knowledge and Data Engineering
1995 – 2002	IEEE Transactions on Pattern Analysis and Machine Intelligence
1995 – 2002	IEEE Transactions on Software Engineering

Table 1: Journals

errors have been found within the data that did not allow some records to be linked correctly.

One surname element for example contained *Hudak at al.*, another one the string *Agrawaland* for an author named *Agrawal*, obviously containing the *and* from the list of authors. Some publications did not have an author specified at all. Altogether 10,382 of the bibliography entries do not have an author specified and got a key using the string 'UNKNOWN' as the author's last name.

Titles have not always been cited correctly, e.g.

```
3 Weighted Pseudo Random Test Generation
3 Weight Pseudo Random Test Generation
```

and 5,136 entries (3.4%) have been skipped altogether, since no title is defined.

Stemming or even probabilistic record linkage techniques could be used to increase the number of correctly identified publications but this potentially introduces false positives and for the scope of this prototype we accepted that some links would be missed, even though this can cause duplicate entries in the reference lists generated by our prototype.

3.3 Retrieval engine

In our experiments, we used the PADRE retrieval system [8]. For text ranking PADRE uses a marginally modified Okapi BM25 relevance function developed by Robertson et al. [13]. PADRE makes use of anchortext extracted from web documents as they are indexed and is also capable of using externally derived anchortext files. Anchortext scoring uses the AF1 formula described by Hawking et al. in [9].

In a second processing iteration we extracted all elements containing one or more references within the text for each journal article.

In addition to using the whole parent element of a reference ($< ref >$) element – usually a paragraph ($< p >$) – as anchortext, we extracted the sentence containing the reference for an alternative index. Sentences are defined as full stop delimited areas around a reference. From these extracted elements we created anchortext records for each reference. In each record the link target was the key we assigned to the reference and the anchortext was either the surrounding sentence or the surrounding paragraph.

We decided to build three different indices using different sources of anchortext:

- AtN - Anchortext not used
- AtS - Anchortext based on Sentences
- AtP - Anchortext based on Paragraphs

The first used no anchortext and the others used sentence-level and paragraph-level anchortext respectively.

To increase precision at search time, and to reduce the length of reference lists to be judged, we configured our search engine to only display full matches. In some cases this caused generated lists to contain only a few or even no records.

Table 2 gives some details of the data indexed.

4 Experiments

In our evaluation the quality of reference lists generated from the three indices was evaluated by fif-

	Quantity	Size (MB)	Size of tar (MB)
Articles	12,107	536	
BibEntries	149,168		
BibEntries with unknown author	10,382		
BibEntries used article title	121,971		
BibEntries used publication title	22,061		
BibEntries skipped (no title)	5,136		
BibEntries files created	96,491	476	140
BibEntries files reused	47,541		
References	241,228		
References without a refid	11		
References found its BibEntry	233,602		
References found no BibEntry	7,615		
Anchortext file paragraph		205	
Anchortext file sentence		85	

Table 2: Preprocessing quantity structure

teen experimental subjects (all researchers from our institution who volunteered to participate) using a comparative approach. The tool used to do the comparison was based on one described by Thomas and Hawing in [15].

After consenting to participate and logging in the subjects were given the task to generate bibliographies by entering topics and to judge the reference lists returned. The subjects were presented with an interface including a search box and were encouraged to enter a query representing a research topic in which they were interested. In response, the comparison tool presented three results lists generated by processing the query against each of the three indices. As can be seen in Figure 2, results were presented in normal bibliography style. The three lists were displayed side-by-side in random order for each query to avoid a bias for or against particular screen locations. Users were able to choose the length of results list before searching and also to request a longer or shorter list at any time during the process. For each of the three bibliographies, the users were asked to assess quality on a scale of 0 to 9 (0/useless - 9/excellent). With each query they were also invited to store a comment on the judging.

All judgments and comments were recorded for analysis.

5 Results

To compare the three versions, we used AtN as our baseline system and compared the subject's judgements relative to that baseline.

The quality of the reference lists generated subjectively varies for different queries and the subjects were allowed to pick their own topics.

User	Judgements		
	AtN	AtS	AtP
User A	1	5	6
User B	0	8	7
User C	2	8	7

Table 3: Different judgements for query 'haptics'

Table 3 shows that subjects judged the results for coincidentally identical queries differently, not only in absolute figures but also relative to each other.

A statistical analysis of variance (ANOVA) of the collected data shows that the participating research scientists preferred the anchortext versions to the plain bibliography entry index with statistical significance with a p-value $\leq .001$ using Fisher's Least Significant Difference (LSD) test (using 242 degrees of freedom). However, no significant preference between the sentence and the paragraph based approach could be identified.

Figure 4 shows the total number of cases, in which the subjects preferred AtP to AtN. Subjects often ranked AtP equally to the baseline, but when they perceived a difference it was most often in favour of AtP. The same comment is valid for Figure 5. AtS is preferred to the baseline.

Figure 6 shows the comparison between AtP and AtS. The values are normal distributed around zero, which visualized the results shown above; the difference between AtS and AtP is not significant.

Figure 3 shows for each subject the mean preference for the sentence or paragraph based version versus the baseline that uses no anchortext at all. In all but one case the anchortext approaches were preferred. Only subject 9 found the list generated

using a paragraph based anchortext approach on average worse (by one point) than the one generated by the baseline system.

6 Discussion

The three different lists presented show the typical tradeoff between precision and recall. The baseline version always exactly matched against the bibliography entry, while the sentence based anchortext approach increased recall at the cost of precision. Some searchers observed, that some entries returned by AtS and AtP do not contain the search term.

To increase the precision we configured our search engine to only display full matches. This caused some of the generated lists to only contain a few or even no records.

The paragraph extracted is defined as the super element of the text containing the reference. Sometimes this super element was not a paragraph element, but for example a table data element. Sentences are defined as full stop delimited areas around a reference.

In the current version, the complete sentences and paragraphs containing multiple references are indexed for each of the references. This might be the source of wrong mappings in cases where a single sentence or paragraph refers to different topics. Developing more sophisticated algorithms to split paragraphs and sentences into units of text more relevant to the referenced publication is expected to increase quality.

Depending on the purpose of the reference list created, different precision/recall tradeoffs might be preferred.

Reference lists generated by our system are presented in order of descending scores assigned by the retrieval system. Anchortext ranking tends to mean that highly cited items will be highly ranked, which is probably an advantage. However, researchers may prefer a date or author ordered listing.

Our subjects didn't always look at the entire reference lists when making their ratings, suggesting that they made their judgments based on early precision or on the presence or absence of expected key items at the head of the list. This implies that the ranking of references is important and that in future work, it would be worth paying attention to optimizing the ranking function for this specialised purpose.

6.1 Alternative evaluation approaches

Our subjects were able to judge the quality of the returned lists on a purely subjective basis. This has its advantages but in future work, we will consider both asking subjects to judge the complete reference lists and asking them to additionally rate the value of each bibliography item. We envisage modifying the three-panel comparison tool to allow such judgments and to set a background colour for each judged item wherever it appears in each list.

As an alternative evaluation approach, a comparison between a survey paper[3]'s bibliography and a reference list automatically generated using our methods was considered. However, this evaluation technique was not selected for fear of restrictively limited overlap between the selected survey paper's bibliographies and papers cited by the IEEE articles in the matching time frame. This method may be able to provide a more absolute type of judgment and will also be considered in future work.

6.2 Biases

The subjects in the pilot experiment were mainly research scientists from within our institution. A different set of subjects might judge the quality of the lists differently. However, it should be noted that the target group of a tool generating reference lists is not the general public.

The nature of the approach and the age of the data set mean that work published after the most recent articles in the INEX collection cannot possibly be retrieved by our system. The only solution to this seems to be to endeavour to obtain more recent data. The tendency for citation counts to increase with the passage of time since publication also means that our anchortext rankings are likely to rank older items higher in the list.

7 Conclusions

Using a three-way side-by-side comparison of reference lists automatically generated from the INEX collection, we have shown that anchortext techniques from web retrieval are also beneficial in the XML retrieval domain. As yet, it is not clear whether the anchortext scope should be at the sentence or paragraph level. Ideally, natural language processing techniques might be used to set the appropriate context for each reference, particularly where multiple references occur within the same paragraph or sentence.

Our tool for generating reference lists from a collection of scientific articles illustrates a useful application for the retrieval of elements other than full documents, and an application that retrieves XML elements based on data outside of that element or its sub tree.

Many avenues have been identified for possible future work, including investigation of alternative evaluation methods, better anchortext extraction and the use of more extensive and up-to-date data. It would be very interesting to study when and

[3]e.g. those in ACM Computing Surveys.

how searchers would actually use a bibliography generation tool if they had access to one, and to perform more thorough evaluation of the tool in the context of real bibliography generation tasks.

Acknowledgements

We want to thank Anne-Marie Vercoustre of INRIA for various discussions and feedback and Alec Zwart of CSIRO Mathematical and Information Sciences for statistical analyses. Finally we'd like to thank the participants in this experiment for their time and efforts.

References

[1] Kurt Bollacker, Steve Lawrence and C. Lee Giles. CiteSeer: An autonomous web agent for automatic retrieval and identification of interesting publications. In Katia P. Sycara and Michael Wooldridge (editors), *Proceedings of the Second International Conference on Autonomous Agents*, pages 116–123, New York, 1998. ACM Press.

[2] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of WWW7*, pages 107–117, 1998. `www7.scu.edu.au/programme/fullpapers/1921/com1921.htm`.

[3] Nick Craswell, David Hawking and Stephen Robertson. Effective site finding using link anchor information. In *Proceedings of ACM SIGIR 2001*, pages 250–257, New Orleans, 2001. `www.ted.cmis.csiro.au/nickc/pubs/sigir01.pdf`.

[4] B. Davison. Topical locality in the web. In *Proceedings of ACM SIGIR'2000*, pages 272–279, Athens, Greece, 2000. `www.cs.rutgers.edu/~davison/pubs/2000/sigir/`.

[5] Norbert Fuhr, Norbert Gövert, Gabriella Kazai and Mounia Lalmas. INEX: INitiative for the Evaluation of XML Retrieval. In *Proceedings of the SIGIR 2002 Workshop on XML and IR*, 2002.

[6] E. Garfield. Citation analysis as a tool in journal evaluation. *Science*, Volume 178, pages 471–479, 1972.

[7] N. Gilbert. A simulation of the structure of academic science. *Sociological Research Online*, Volume 2, Number 2, 1997.

[8] David Hawking, Peter Bailey and Nick Craswell. Efficient and flexible search using text and metadata. Technical Report TR2000-83, CSIRO Mathematical and Information Sciences, 2000. `http://www.ted.cmis.csiro.au/~dave/TR2000-83.ps.gz`.

[9] David Hawking, Trystan Upstill and Nick Craswell. Towards better weighting of anchors (poster). In *Proceedings of SIGIR'2004*, pages 99–150, Sheffield, England, July 2004. `http://es.csiro.au/pubs/hawking_sigirposter04.pdf`.

[10] R. Larson. Bibliometrics of the world wide web: An exploratory a nalysis of the intellectual structure of cyberspace. In *Annual Meeting of the American Society Information Sc ience*, 1996.

[11] Huajing Li, Isaac Councill, Wang-Chien Lee and C. Lee Giles. Citeseerx: an architecture and web service design for an academic document search engine. In *Proceedings of WWW 2006, May 2326, 2006, Edinburgh, Scotland.*, 2006.

[12] Oliver A. McBryan. GENVL and WWWW: Tools for Taming the Web. In *ProceeProceedings of the First International World Wide Web Conference 1994*, 1994.

[13] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu and M. Gatford. Okapi at TREC-3. In D. K. Harman (editor), *Proceedings of TREC-3*, Gaithersburg MD, November 1994. NIST special publication 500-225.

[14] H. Small. Co-citation in the scientific literature: A new measu re of the relationship between two documents. *Journal of the American Society for Information Scien ce*, Volume 24, 1973.

[15] Paul Thomas and David Hawking. Evaluation by comparing result sets in context. In *To appear in Proc. CIKM 2006*, 2006.

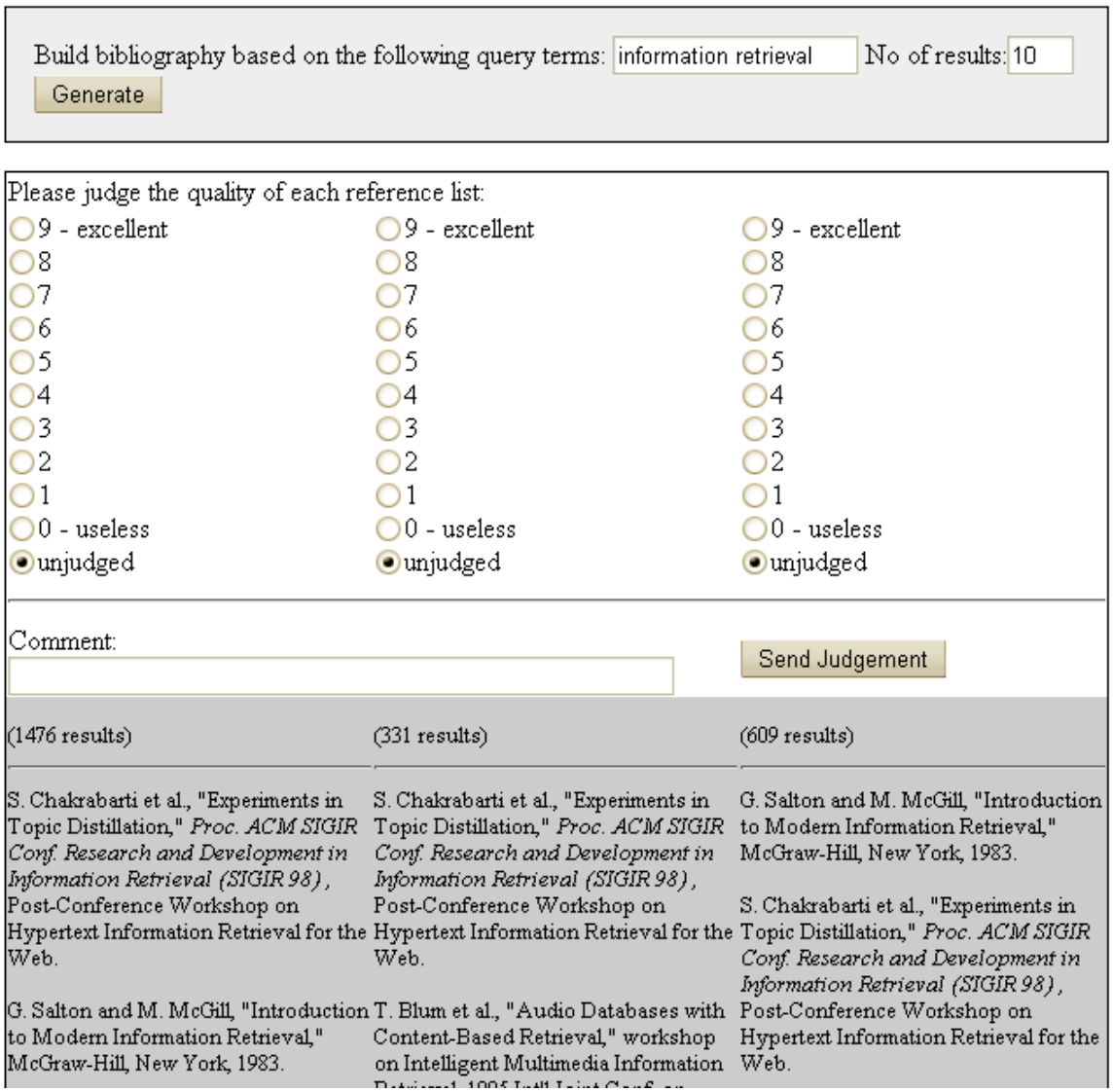

Figure 2: Judging interface

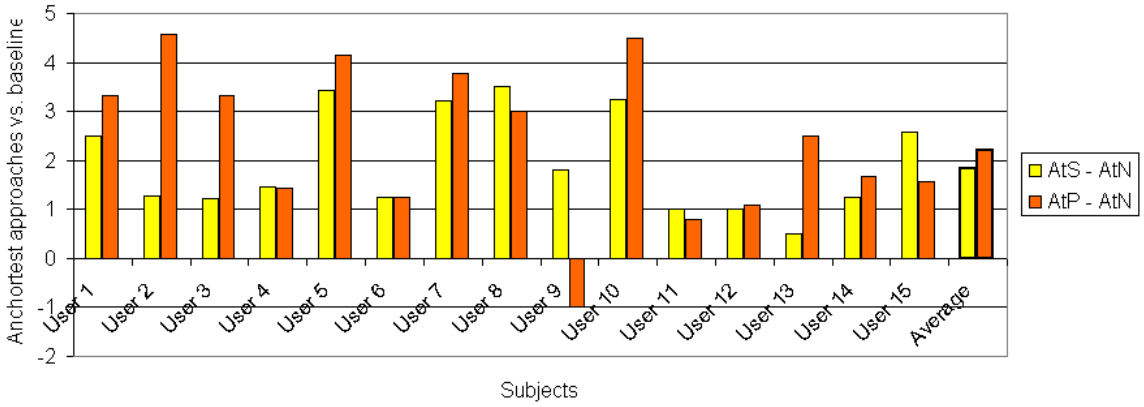

Figure 3: User Comparison

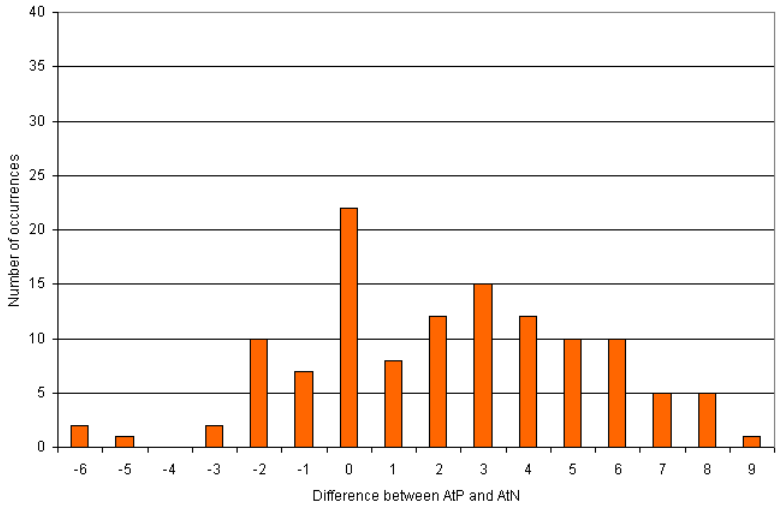

Figure 4: Distribution of AtP rating minus AtN rating

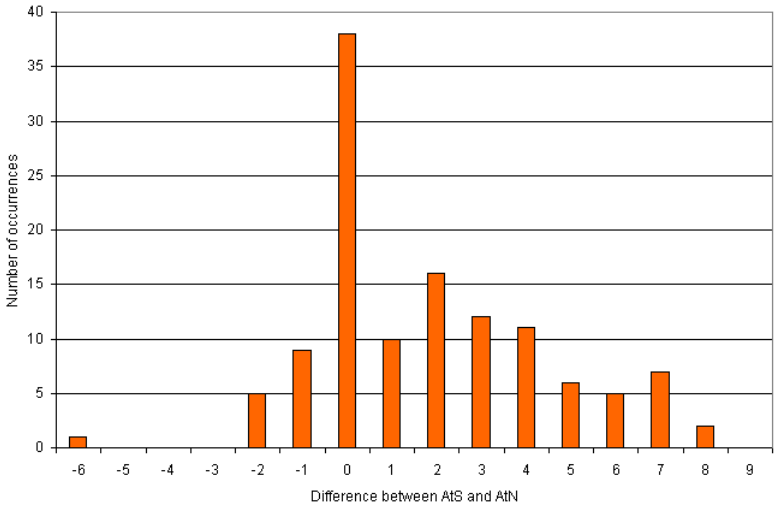

Figure 5: Distribution of AtS rating minus AtN rating

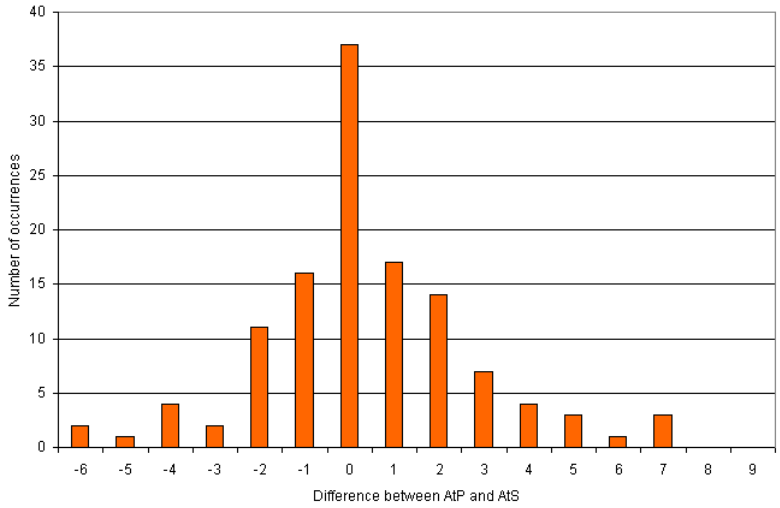

Figure 6: Distribution of AtP rating minus AtS rating