Construction of an Annotation Test Collection for Search Evaluation on Digital Libraries

Maristella Agosti* Tullio Coppotelli* and Nicola Ferro*

*Department of Information Engineering
University of Padua – Italy
{agosti, coppotel, ferro}@dei.unipd.it

***Abstract*— This study presents an automatic method for constructing an experimental test collection with annotated documents. This collection can be used to evaluate the search functionalities of digital library systems that manage both documents and annotations. The method makes use of an existing collection composed of two sets of documents. Starting from the relevance assessments of the original collection a graph is built that contains enough information for the identification of the best links between documents and annotations. Using these links the annotation collection is built.**

I. Introduction

Annotation is a concept we all are familiar with, since it is common to take notes while writing or reading a document. Annotations are not limited to the paper form. Instead, we can take full advantage of them by providing a Digital Library System (DLS) with annotation capabilities on its digital documents. Annotations make it possible to expand the contents of a Digital Library with personal information resources and integrate them into the users' way of working.

Several studies have been performed to identify newer and better algorithms which aim to improve retrieval effectiveness and better satisfy end-users information need. In this perspective, the annotations made on documents offer an interesting possibility for improving information access performances. Indeed, the additional information contained in the annotations and the hypertext which connects annotations to documents allows us to define search strategies which merge multiple sources of evidence to increase system effectiveness.

Methods are needed which allow us to search for annotated documents by exploiting the annotations linked to them. The aim of this kind of search is to retrieve better and more documents with respect to a search without using annotations.

Golovchinsky et al. [1] proposed the use of annotations manually written by users over electronic documents, as a way to implement query expansion and relevance feedback. The results showed how this approach increases the effectiveness of the system with respect to the simple use of relevance feedback, but it is limited to only one facet: the use of annotation as a relevance feedback. Frommholz et al. [2] proposed a system that implements annotations for collaboration among scholars. In that system annotations were used to provide advanced content and content-based access to the underlying digital repository. This work adopted a broader view annotations and it enables the creation of a collaborative experience over the Digital Library (increasing the user experience of the DLS) but it does not present any evaluation of the system effectiveness. Agosti and Ferro [3] proposed an algorithm that allows the concurrent search of documents over different DL using annotations. Annotations were used to naturally merge and link personal contents with the information resources provided by the DLS and were exploited during the research not only to rank documents better but also to retrieve more relevant documents.

However, while different approaches based on annotations have been proposed, they still lack a full experimental evaluation. This evaluation is needed to investigate the effectiveness of this kind of search and to assess whether the performance improves with respect to a search without making use of annotations. When it comes to the experimental evaluation of search algorithms which exploit annotations, there is the problem of the lack of test collections that include annotations. Therefore, this study addresses this lack and presents a method for the creation of such a test collection, which has been initially proposed in [4].

II. A New Approach to Collection Construction

As explained above, there is a need for a test collection for assessing a specific task: the retrieval of documents exploiting annotations. Usually there is no other choice but to create the new collection from scratch. In this case, we propose an alternative solution that involves the use of an already available test collection and the automatic construction of a parallel collection of related annotations. This method has three main advantages: 1) it reduces the effort of creating the new collection, since the collection construction is completely automatic; 2) the results obtained with the newly created collection are comparable with the previous results obtained on the original test collection; 3) the approach exploits existing pools to deal with a sufficient number of experiments; 4) it is not limited to the creation of only one collection but it allows the creation of multiple collection with different characteristics.

Particular attention has to be paid in checking the quality of the new test collection, since it is very important that the experimental results obtained are as reliable as those that would have been obtained using a new collection built with human effort. To address this aspect, it is important to understand which kind of relationship is established between a document and an annotation: for example, annotations can

expand some aspect of the annotated document or propose different points of view. In this context, it is worth pointing out that those kinds of relationship can be automatically created among documents of existing collections. Imagine a collection of journal articles: articles written by different authors about the same subject (different points of view) can be found in such a collection or articles written by the same author but on different days (expansion of the original contents) are present. The aim of our method is to determine those relationships and to create a new collection where the documents that can be considered as annotations are found and linked to the documents with which they hold those specific relationships. The main idea is to use relevance assessments, produced by human assessors, as the initial base to consider the documents as annotations.

The following present an intuitive explanation of the proposed method. A thorough formalization has been done and is available in [5].

The proposed method makes it necessary to find a collection that is naturally divided in at least two sets of independent documents, with similar informative contents so that one is the set of documents and the other can be considered as a set of candidate annotations, e.g. a collection of journal articles from the same newspaper written over the same period of time. Starting from the relevance assessments, we create a graph that enables us to have a comprehensive view of relationships between documents. This graph has three kinds of vertices: topics T, documents D, and candidate annotations A. Each positive relevance assessment allows us to create an edge (i, j) between a topic T_i and either a document D_j or a candidate annotation A_j. These edges connect candidate annotations to documents using human relevance assessments and passing through topics. The proven quality of these paths in the graph is fundamental to our method.

The presence of cycles in the graph allows us to strengthen the relationship between documents and annotations because the path between a document and an annotation is supported by a greater number of human relevance assessments.

The proposed algorithm tries to annotate as many documents as possible in the collection, bearing in mind those two situations. First, it identifies all couples of documents belonging to at least one cycle in the graph. Second, the algorithm resolves conflicts, like that of figure 1a, where if the annotation A_1 had assigned to document D_2, it would not have been possible to annotate document D_1. Having dealt with both these positive and negative situations, all the paths left in the graph have the same importance and it is now possible to proceed to resolving all other conflicts with random choices.

It has been showed in [4], that the proposed algorithm identifies, in an experimental context, the greatest possible number of annotated documents. It appear that, although a counter example in the general case can be found, the algorithm reaches exactly the upper boundary when applied to real collections.

We have hints that in the original collection more document-annotation relationships exist than those that this first method can identify. Therefore, we decided to find an alternative strategy to integrate the annotations already found, as discussed in the following section.

III. Subtopic as a Way to Expand the Graph

Since very few documents are usually relevant for each topic, we decided to expand the graph to make use of the whole pool instead of only the positive relevance assessments. We proceed like in the previous method, creating edges between documents and annotation but, having lost the support of human assessments, the paths in the graph may no longer indicate a strong relationship between a document and an annotation and so it now becomes necessary to estimate the strength of a relationship. By using the whole pool, we aim to identify relationships between those documents which are not relevant to any specific topic but instead are good annotations for other documents in the collection. In fact, considering a single topic and examining the retrieved documents (both relevant and not relevant), we noticed that these documents can be separated in different sets on the basis of their arguments creating new topics called subtopics (S_i, as illustrated in figure 1b). As a consequence, there are documents in the graph that are not relevant to the original topic but are relevant to some specific subtopic. An example of this statement can be found in a topic about the skiing victory of Alberto Tomba: the method proposed in Section II using only documents relevant to this topic can annotate few documents, while the method proposed in this section can find other good annotations using all documents relevant to subtopics S_i like general skiing reports or documents about Tomba's public life or problems with the law.

Weighting each edge in the graph, we aim at identifying new relationships between documents and annotations that belong to the same subtopic. Without reassessing the documents of the collection on the basis of these new (sub-)topics, we propose to use four parameters, automatically calculated from the graph and the contents of the document: 1) affinity; 2) a score obtained using an information retrieval tool; 3) generality; and 4) temporal proximity.

Firstly, the affinity is the normalized number of cycles that two or more topics have in common and is a measure of the superimposition of the content of the involved topics. We noticed that the greater the affinity P_A, the greater is the probability of finding document-annotation pairs of good quality. Moreover, we noticed that pairs with minimum affinity more probably result from random coupling rather than from real content superimposition and the edges between these topics can be discarded.

Secondly, we index the entire set of candidate annotations and we build a query using the contents of documents to assess the similarity between documents and candidate annotations. The results above a threshold are intersected with the graph: all the paths outside the intersection are removed from the graph, while all the other paths are weighted with weight P_{IR}.

Thirdly, the generality is the number of topics for each document-annotation pair. The lower the generality, the higher is the score P_G assigned to a document-annotation pair.

Finally, the score P_T reflects the temporal proximity of two documents. We noticed that the quality of a document-annotation pair increases when an annotation is temporally

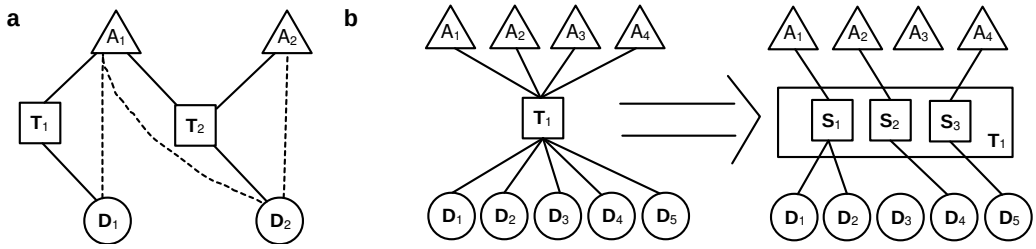

Fig. 1. Conflict (a) and subtopic (b) in the graph.

close to a document and finding valid pairs between documents written within a few days each other becomes more likely.

Given these parameters, we propose a formula that allows us to obtain a unique score with which we can weight all the remaining paths in the graph: $P = \alpha_A * P_A + \alpha_{IR} * P_{IR} + \alpha_G * P_G + \alpha_T * P_T$ where $\alpha_A + \alpha_{IR} + \alpha_G + \alpha_T = 1$. The score P is in the range $[0, 1]$, where 1 is the best document-annotation pair.

Given this score it is possible to distinguish between good and bad document-annotation couples. Discarding the worst ones and selecting the best ones more documents of the original collection are annotated. There is a trade-off between the number of new documents that we can annotate and the quality of the selected annotations and, for this reason, we select only annotations that can annotate new documents without decreasing the overall quality of the newly created collection.

IV. Conclusions and Future Work

The proposed algorithm enables the creation of an experimental test collection where documents are annotated with an adequate number of annotations of good quality. Since the annotations are obtained by simulating human behavior, the process improves the collection reliability. Moreover, this approach is not limited to the creation of a single annotated test-collection but, starting from different collections, different results can be obtained (e.g. collection with monolingual or multilingual documents).

The affinity parameter has been proposed and introduced here for the first time. Apart from its use in the context of test collection building, we are interested in verifying the possibility of using it in other contexts and models. A possibility is to investigate if the proposed method can be used in helping to reduce the number of documents that assessors have to judge in traditional evaluation activities and to find a strategy that allows us to evaluate a posteriori the relevance assessments.

The next step will be the set up of an adequate test bed and the consequent evaluation of existing algorithms that use the annotations with the provided collection to compare the obtained results with these obtained by algorithms that do not use annotations. This comparison will provide good information for both the correctness of our collection and the efficacy of these algorithms.

Acknowledgements

The work reported in this paper has been partially supported by the DELOS Network of Excellence on Digital Libraries, as part of the Information Society Technologies (IST) Program of the European Commission (Contract G038-507618).

References

[1] G. Golovchinsky, M. N. Price, and B. N. Schilit, "From Reading to Retrieval: Freeform Ink Annotations as Queries," in *Proc. 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1999)*, F. Gey, M. Hearst, and R. Tong, Eds. ACM Press, New York, USA, 1999, pp. 19–25.

[2] I. Frommholz, H. Brocks, U. Thiel, E. Neuhold, L. Iannone, G. Semeraro, M. Berardi, and M. Ceci, "Document-Centered Collaboration for Scholars in the Humanities – The COLLATE System," in *Proc. 7th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2003)*, T. Koch and I. T. Sølvberg, Eds. Lecture Notes in Computer Science (LNCS) 2769, Springer, Heidelberg, Germany, 2003, pp. 434–445.

[3] M. Agosti and N. Ferro, "Annotations as Context for Searching Documents," in *Proc. 5th International Conference on Conceptions of Library and Information Science – Context: nature, impact and role*, F. Crestani and I. Ruthven, Eds. Lecture Notes in Computer Science (LNCS) 3507, Springer, Heidelberg, Germany, 2005, pp. 155–170.

[4] T. Coppotelli, "Creazione di una collezione sperimentale per la valutazione di sistemi di reperimento dell'informazione che utilizzino le annotazioni (in Italian)," Master's thesis, Department of Information Engineering, University of Padua, 2006.

[5] M. Agosti, T. Coppotelli, N. Ferro, and L. Pretto, "Exploiting Relevance Assessment for the Creation of an Experimental Test Collection to Evaluate Systems that Use Annotations," in *DELOS Conference 2007 Working Notes*, C. Thanos and F. Borri, Eds. ISTI-CNR, Gruppo ALI, Pisa, Italy, February 2007, pp. 195–202.