An Approach for the Construction of an Experimental Test Collection to Evaluate Search Systems that Exploit Annotations

Maristella Agosti, Tullio Coppotelli, Nicola Ferro, and Luca Pretto

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

Abstract. This study addresses the lack of an adequate test collection that can be used to evaluate search systems that exploit annotations to increase the retrieval effectiveness of an information search tool. In particular, a new approach is proposed that enables the automatic creation of multiple test collections without human effort. This approach takes advantage of the human relevance assessments contained in an already existing test collection and it introduces content-level annotations in that collection.

1 Introduction

The topic of annotations is focusing researchers' attention in both the *Digital Library (DL)* and *Information Retrieval (IR)* fields. In DLs, annotations are used to facilitate cooperation between users [1], to enrich the content of documents or to easily describe documents in media different from plain text, like video or audio. In IR, new and better algorithms which aim to improve the system retrieval effectiveness using annotations has been proposed. In fact, annotations offer an interesting opportunity to improve the retrieval performance: the additional information contained in the annotations and the hypertext which connects annotations to documents enable the definition of search strategies which merge multiple sources of evidence in order to increase the system effectiveness. In this perspective, two approaches have been proposed in [2] and [3]. The former presents a theoretical model and discusses how it exploits annotations and the hypertext that documents and annotations constitute in the retrieval process. The latter exploits annotations as a rich source of evidence to augment the content of each document with the content of its attached annotations.

The evaluation of the effectiveness of these approaches is a necessary step that enables not only understanding of their effective performance but also, at a more general level, confirmation that annotations can play an important role in improving system effectiveness. In [2] the authors stressed that an obstacle to the complete evaluation of these kinds of systems was the lack of an experimental test collection. In [3] an effort in this direction has been made with the manual creation of a small test collection that is used to evaluate their own approach.

C. Thanos, F. Borri, and L. Candela (Eds.): Digital Libraries: R&D, LNCS 4877, pp. 167–176, 2007.

The greatest difficulty encountered by the author during the creation process was the need to cope with limited resources for the relevance judgement creation.

The creation of an experimental test collection is a consolidated process in IR and an overview of this process and related problems is given in [4]. Despite this, when it comes to the creation of a test collection with annotated documents, the problems which need to be addressed are demanding. To summarize, the creation of a test collection with annotated documents requires the finding of a suitable set of documents that have to satisfy required characteristics, the manual creation of the annotations over these documents, the creation of the topics containing the information that have to be searched and, finally, the evaluation of the document relevance to each topic. Moreover, as we will discuss in greater detail in Section 2, different views of the annotations can lead to the need for different test collections; therefore a different approach to test collections creation would be suitable to cope with this aspect. Instead of using the limited resources to obtain human relevance assessments, an entirely automatic technique can be envisaged. Therefore, the problem of setting an adequate experimental test-bed for search algorithms which exploit annotations was addressed. A flexible strategy to create test collections with annotated documents was identified that, starting from an already existing test collection, brings to the surface the hidden work made by the assessors during the creation of relevance assessments. An interesting feature of this strategy is that it is not limited to the creation of a single test collection, rather by using as a starting point collections with different characteristics, it allows the creation of new collections with the same characteristics as the original one (monolingual or multilingual, general or specialized). An initial proposal was made in [5] and the progress was reported in [6].

The final aim of the research is to establish a framework reusable for the evaluation of different information search tools. This paper reports on the proposed approach and the so called *subtopic view of the graph* that is the starting point for a new algorithm, proposed in [7], which enables the construction of a less sparsely annotated test collection that could be used to evaluate *Information Retrieval Systems(IRSs)* under different testing conditions. This paper and [7] can be considered complementary as they both report on the general approach of constructing a test collection to be used for evaluating search tools that use annotations together with the original documents to solve users information needs. To reach the objective of the study, Section 2 presents an introductory overview on the annotation concepts. Section 3 reports on the adopted approach and Section 4 uses the approach to describe an algorithm that exploits relevance assessments to introduce annotations in the original test collection. Section 5 discusses the obtained test collection. Conclusions and future work are presented in Section 6.

2 Overview on Annotations

The concept of annotation is not limited to the action of a scholar who annotates a text passage writing annotations as it is a rather more complex and multifaceted concept; the concept has been addressed at length and an extensive study is [8].

An intrinsic dualism exists in annotations. Are annotations a *content enrichment* or are they *stand-alone documents*? If they simply enrich the content of the original document, then the annotations can be merged to that document and then annotation and document become an atomic document. The approach to the test collection creation described in [3] adopted this view. As a result, annotations are not autonomous entities, but rely on previously existing information resources to justify their existence. The approach to retrieval adopted in [2] considers the annotations as autonomous objects. Stand-alone annotations can be evaluated relevant to a topic regardless to the document relevance. These two opposite approaches stress how broad the concept of annotation can be and, as a consequence, how hard it can be to build a test collection that enables the evaluation of systems that use annotations.

An important characteristic of annotations is their heterogeneity. Annotations over documents can be created at different times and by different authors each with a different background. The user who annotates a document may know recent information about the topic that the document author did not know. He may disagree with the document content and might like to communicate his different opinion. The author of the document can clarify or modify some text passage. This heterogeneity is a key-point that allows a dynamic improvement in the content of the document and by using this new information it is possible to better estimate the relationship between documents and query, a feature which is so important in document retrieval.

Summing up, the goal of the approach presented in the following Section is to enable the creation of test collections that respect the different aspects of annotations without the need for extensive human effort.

3 The Proposed Approach

3.1 Overview

When building a test collection from scratch finding a suitable set of documents, creating the topics and evaluating the document relevance to each topic are required. All these tasks are not trivial [4] and need an accurate evaluation to avoid the introduction of too many biases in the test collection. The task that we are going to address is even more difficult. The creation of annotations over the selected documents is a very expensive practice. Moreover, it is not possible to use assessors to create the annotations because, to maintain their heterogeneous nature, a wide range of annotations written by different authors in different periods of time would be needed. The pooling method is a consolidated practice that reduces assessor effort during the creation of the relevance judgments. This method requires the assessment of only a reduced number of document for each topic – i.e. 1000. The pooling method relies on a certain number of experiments that are performed with different *Information Retrieval Systems (IRSs)* but the number of systems that use annotations is currently too small to allow the creation of a sufficient number of experiments and this prevents us from using this method. Finally, if we were able to overcome these limitations the standard collection creation process would still be expensive and time consuming.

Our approach avoids all these problems and proposes a different strategy that involves the use of an already existing test collection as a starting point and the automatic construction of a parallel collection of related annotations. This strategy has the following advantages:

1. it reduces the overall effort needed to create the test collection;
2. the results obtained evaluating systems with the new collection are comparable with the previous results obtained on the original test collection; this allows the direct performance comparison between systems that use annotations and systems that do not use them;
3. it exploits the existing pool to deal with a sufficient number of experiments;
4. it allows the creation of multiple collections with different characteristics and the consequent evaluation of the system behavior in different contexts.

The idea is to use as a starting point a test collection that contains documents that are naturally separated in two or more different sets and to use the relevance judgments of the human assessors to link documents that belong to different sets. If document d_i and document $\hat{a}_j$ were both judged relevant to the topic t_z by a human assessor then we know that these documents are put in relation by the content of the topic. We then use the topic as the motivation for the document $\hat{a}_j$ to annotate the document d_i. In this way a set of annotated documents can be created whose relevance to the topics has already been judged in the original test collection.

3.2 The Modelling

The starting test collection can be represented as a triple $C = (D, T, J)$ where D is the set of documents, T is the set of topics and J is the set of relevance assessments defined as $J = D \times T \times \{0, 1\}$ (binary relevance). The documents D of the chosen test collection must be divisible in two disjoint sets, D_1 and $\hat{A}$, where $D = D_1 \cup \hat{A}$ and $D_1 \cap \hat{A} = \varnothing$. We have conducted preliminary experiments where D_1 were newspaper articles and $\hat{A}$ were agency news of the same year [5]. The annotated collection is $C' = (D'_1, T, J)$, where D'_1 contains exactly the same documents as D_1 with the addition of annotations over these documents. Topics and relevance assessments are exactly the same. In C' we use a subset A of $\hat{A}$ to annotate the documents in D_1, thus Â is the set of candidate annotations and A is the set of actual annotations. The goal is then to find which candidate annotations can be used to correctly annotate documents in D_1 and create the annotation hypertext over these documents. To identify these relationships we take advantage of the fact that in C the topics are made over both D_1 and $\hat{A}$ (thus their relevance to each topic has been judged): if in C both a candidate annotation and a document have been judged relevant to the same topic then we infer that it is possible to annotate that document with that candidate annotation. Referring to Figure 2, these couples (document, annotation) are those connected by a two-edge path in the undirected graph $G_1 = (V_1, E_1)$ where $V_1 = D_1 \cup T \cup \hat{A}$ and $E_1 = (D_1 \cup \hat{A}) \times T$. In G_1 each edge

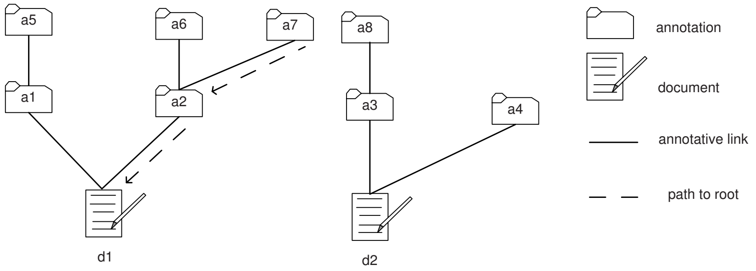

Fig. 1. Annotation constraint

represents a human assessment i.e. a path between annotation $\hat{a}_j$ and document d_i passing through topic t_z means that a person assessed both $\hat{a_j}$ and d_i relevant to t_z. This relevance property creates a path between documents and candidate annotations that is used in Section 4 to introduce annotations in C'. The intuition is that the strength of these paths allows the use of candidate annotations as real annotations for connected documents and that these annotations reflect human annotative behaviour.

The proposed approach respects the so called *annotation constraint*: each annotation can annotate one and only one document or annotation; this means that each annotation is written for exactly one *Digital Object (DO)* but it can still be linked to more DOs [2,8]. This constraint has been introduced to better simulate the annotative behaviour of a user who usually writes an annotation only after the reading of a specific DO. As a consequence of this constraint, the set of documents and annotations become a forest and it is possible, starting from an annotation, to identify the root document of an annotation thread. Consider, for example, Figure 1 where annotation a_7 belongs to the tree rooted in d_1 and note that this would not be possible if a_7 could annotate also a_3. This constraint also has the advantage of allowing the identification for each annotation, independently of its depth, of a single root document.

4 Exploiting the Relevance Assessments to Annotate Documents

Once graph $G_1 = (V_1, E_1)$ is given, the problem of matching a candidate annotation with a suitable document can be addressed. The matches should respect the annotation constraint that one annotation can annotate only one document. This section describes an algorithm which makes use of the positive relevance assessments to match a candidate annotation with a document. The first aim of the algorithm is to match each candidate annotation with the most suitable document. When more than one match is possible, the algorithm heuristically tends to choose matches which maximize the number of annotated documents—indeed, maximizing the number of annotated documents is the second aim of the algorithm.

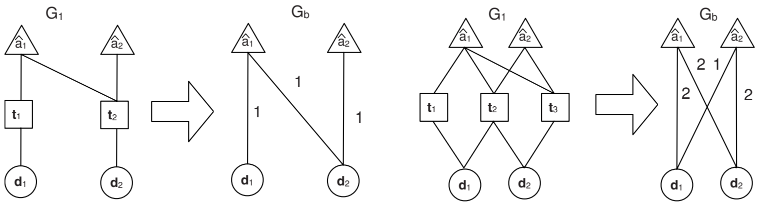

Fig. 2. Examples of the construction of graph G_b, starting from graph G_1

The algorithm works in two phases. In the first phase it constructs a weighted bipartite graph G_b on the basis of G_1, i.e. the graph whose edges represent positive relevance assessments. In the second phase the algorithm works on the weighted bipartite graph G_b to properly match a candidate annotation with a document.

The construction of the weighted bipartite graph $G_b = (V_b, E_b)$ is immediate: the vertices of G_b are all the vertices of G_1 which represent documents or candidate annotations, that is $V_b = D_1 \cup \hat{A}$, and an edge between candidate annotation $\hat{a}$ and document d exists if and only if $\hat{a}$ and d have been judged relevant to at least one common topic, that is $t \in T$ exists such that edges $\hat{a}$-t and t-d are in E_1. Moreover, a weight is assigned to each edge $\hat{a}$-d in E_b, which gives the number of common topics between $\hat{a}$ and d. These weights take account of the fact that when $\hat{a}$ and d are assessed as relevant to more than one common topic at the same time, it is reasonable to suppose that the bond between the candidate annotation $\hat{a}$ and the document d will be strengthened. In Figure 2 simple examples of the construction of G_b, starting from G_1, are given.

Once G_b is constructed, the algorithm works only on G_b to properly match a candidate annotation with a document. It is this second phase of the algorithm that has the two aims described above. The first aim is that of matching the best possible annotation with a document: this is done considering first the edges with the highest weight. The second aim is that of trying to annotate the maximum number of documents, once the best possible annotations have been considered.

The first aim is achieved by first analysing only the edges with the maximum weight and using all of them to match candidate annotations with their suitable documents. After all the edges with the maximum weight have been analysed, only the edges of immediately lower weight are analysed and so on, until all the edges with a positive weight have been analysed. In other words, the algorithm considers each different layer of edge weight separately—the higher the layer, the higher the quality of the matches. When a layer with a certain weight is considered, only edges with that specific weight are analysed.

The second aim, i.e. trying to annotate the maximum number of documents, is achieved by the conceptual application, layer by layer, of two operators, $O_{\text{conflicts}}$ and O_{random}. The first operator is applied to match a candidate annotation with a document, and also has the task of resolving conflicts like those in Figure 3a, where if $\hat{a}_1$ were matched with d_2 it would no longer be possible to annotate document d_1, while the best choice is to match $\hat{a}_1$ with d_1 and $\hat{a}_2$ to d_2.

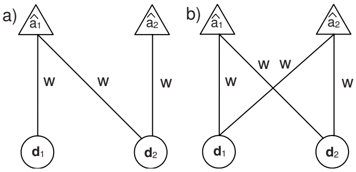

Fig. 3. On the left example of a conflict, on the right example of a deadlock. Note that all edges have the same weight w.

To avoid these conflicts operator $O_{\text{conflicts}}$ first selects all the couples $\hat{a}$-d for which $\hat{a}$ can annotate only one document, like the couple $\hat{a}_2$-d_2 in Figure 3a. Then $O_{\text{conflicts}}$ matches candidate annotations with documents in order to annotate the maximum number of documents: for instance, in the case of Figure 3a, $\hat{a}_1$ will be matched with d_1, since d_2 has already been annotated. Once an edge $\hat{a}$-d is used in a match, it is marked with the negative weight -1, and all the other edges which are incident with the same candidate annotation $\hat{a}$ are deleted from the graph and no longer considered. $O_{\text{conflicts}}$ is iterated until it resolves all possible conflicts. However, in some cases $O_{\text{conflicts}}$ cannot find a match, since no preferable match is suggested by the topology of the graph. This occurs, for instance, when a kind of deadlock exists (see Figure 3b).

In this case an operator O_{random} is applied, which randomly selects one of the possible matches between a candidate annotation and a document. As usual, when a match, that is an edge $\hat{a}$-d, is selected, that edge is marked with the negative weight -1, and all the other edges which are incident with $\hat{a}$ are deleted. The algorithm applies iteratively $O_{\text{conflicts}}$ and O_{random} operators until all the edges with the weight under consideration have been examined. Then a lower weight is examined and so on, until all *positive* weights have been examined.

Finally, edges marked with the negative weight -1 give the desired matches of candidate annotations with documents.

In figure 4 one possible solution of a deadlock problem is proposed. There are four equiprobable edges and if $O_{conflicts}$ cannot match any annotation to document then O_{random} is applied and deletes one edge with probability 0.25. In the example, after the deletion of edge $d_1 - \hat{a}_2$, it is possible to annotate d_2 with annotation $\hat{a}_2$. In the next execution step $O_{conflicts}$ is reapplied that now can match $\hat{a}_1$ with d_1, and not $\hat{a}_1$ with d_2 because d_2 with respect to d_1 is already annotated. Note that by applying O_{random} it is no longer possible to find a unique solution to the matching problem, but this is not relevant with respect to our aim of finding the maximum number of matches.

5 Discussion

The proposed approach is completely automated and allows the creation of a test collection containing a number of documents equal to the cardinality of D and a certain number of annotations over these documents. The number of topics is

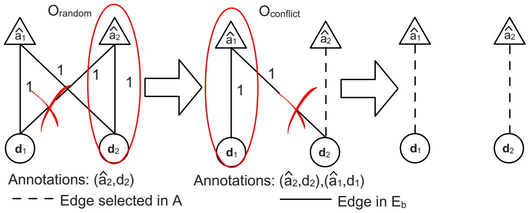

Fig. 4. Example of application of O operators

equal to the cardinality of T while the number of annotations depends on the structure of the graph G_b. Because the graph G_b is build starting from the relevance assessments, it is clear that the number of annotations that this method can introduce strongly depends on the number and distribution of relevance assessments. In this way it is possible to match only the annotations that are assessed relevant to at least one topic. This relationship causes that number to slightly change using different collections and we can state that the test collection obtained with the method presented in Section 4 can be used to simulate a collection with a limited number of annotations with respect to the number of documents. From the point of view of the evaluation, this result is already a good starting point that should enable an initial evaluation of the change in effectiveness of IRSs that use annotations.

The previous algorithm cannot decide anything about candidate annotations that are still in the pool but are not relevant to any topic, because, for construction, in graph G_1 they are not connected to any topic. Despite this, in the original collection there still exists a certain number of candidate annotations that could be correctly used to annotate documents in D. This Section presents a practical justification for their existence. The idea is to build the graph using not only the relevance assessments but also all the information contained in the original test collection, like the information on the documents that entered the pool, the content of both documents and annotations, and, if they exist, metadata about the documents.

We define A_2 as the set of effective annotations identified with the previous algorithm and E_2 as the edges incident to A_2. We define $G_2 = G(V/A_2, E/E_2)$ where G_2 is the graph obtained using the whole pool for each topic in the original collection and removing, due to the annotation constraint, all the candidate annotations already matched. In this new graph we have, for each topic, a set of documents and annotations that are no longer judged relevant to the topic, since those relevant were already assigned by the previous algorithm, but that are still valuable. Consider the graph that represents all documents of G_2 inserted in the pool for a single topic. It is possible to group the vertex of the graph in subsets on the basis of the document contents. For each subset a new topic is ideally created. These new topics are called *subtopic* $S_1, S_2, \ldots, S_k$. The attention is no longer focused on the relevance of documents or annotations to the original topic – since we know by construction that it is not possible – but is focused on finding the couples

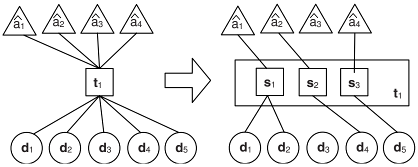

Fig. 5. Subtopic creation

(document, annotation) that are somehow relevant to the same subtopic, like $(d_1, \hat{a}_1)$ and $(d_2, \hat{a}_1)$ in Figure 5.

The following example from a CLEF[1] collection can help to better understand the concept of subtopic. The original topic t is about "Alberto Tomba's skiing victories". The IRSs used to create the pool with the pooling method introduced in the pool for this topic not only relevant documents but also not relevant ones. These not relevant documents can be grouped on the basis of their content in subtopics like "Documents about skiing competitions where Alberto Tomba does not participates" (S_1), "Documents about skiing competitions where Alberto Tomba participates without winning" (S_2) or 'Documents about the social life of Alberto Tomba" (S_3).

It would be useful to find the documents $d_i \in D, \hat{a}_j \in \hat{A}$ where both d_i and $\hat{a}_j$ are incident to the same subtopic S_k. Then the candidate annotations $\hat{a}_j$ could be used to correctly annotate documents d_i. It is important to stress that the goal is not to identify these subtopics, but to find documents that belong to the same cluster. The existence of these subtopics can be used as a practical justification for the algorithm proposed in [7]. Clearly, it is not trivial to find documents and annotations relevant to the same subtopic without knowing these subtopics and to this aim the previous algorithm cannot be used because it relies on human assessments that simply do not exist for subtopics.

6 Conclusions and Future Work

In this paper we pointed out the lack of adequate test collections as the main cause for the incomplete evaluation of IRSs that use annotations to increase system effectiveness. A new and completely automated approach to the creation of necessary test collections has been proposed. The approach is based on an already existing test collection, without annotations, and automatically adds annotations to the collection documents. The reliability of the created test collection is based on the reliability of relevance assessments made by human assessors and hence it has the same quality as the original test collection. The approach is not confined to the creation of a single test collection, but it can be used to create different test collections with different characteristics. The natural continuation of this work is the evaluation of existing systems with the final aim of understanding whether or not the annotations can play an important role in increasing the IRSs effectiveness.

[1] http://www.clef-campaign.org/

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). The work of Tullio Coppotelli was partially supported by a grant of MicroWave Network S.p.A., Padua, Italy.

References

1. Ioannidis, Y., Maier, D., Abiteboul, S., Buneman, P., Davidson, S., Fox, E.A., Halevy, A., Knoblock, C., Rabitti, F., Schek, H.J., Weikum, G.: Digital library information-technology infrastructures. International Journal on Digital Libraries 5, 266–274 (2005)
2. Agosti, M., Ferro, N.: Annotations as Context for Searching Documents. In: Crestani, F., Ruthven, I. (eds.) CoLIS 2005. LNCS, vol. 3507, pp. 155–170. Springer, Heidelberg (2005)
3. Frommholz, I.: Annotation-based document retrieval with probabilistic logics. In: Fuhr, N., Kovacs, L., Meghini, C. (eds.) ECDL 2007. Proc. 11th European Conference on Research and Advanced Technology for Digital Libraries. LNCS, vol. 4675, pp. 321–332. Springer, Heidelberg (2007)
4. Voorhees, E.M., Harman, D.K. (eds.): TREC: Experiment and Evaluation in Information Retrieval. MIT Press, Cambridge (2005)
5. Coppotelli, T.: 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)
6. Agosti, M., Coppotelli, T., Ferro, N., Pretto, L.: Exploiting relevance assessments for the creation of an experimental test collection to evaluate systems that use annotations. In: DELOS Conference, Pisa, Italy, pp. 195–202 (2007)
7. Agosti, M., Coppotelli, T., Ferro, N., Pretto, L.: Annotations and digital libraries: Designing adequate test-beds. In: Goh, D.H., Cao, T., Sølvberg, I., Rasmussen, E. (eds.) ICADL 2007. Proc. 10th International Conference on Asian Digital Libraries. LNCS, Springer, Heidelberg (in print, 2007)
8. Ferro, N.: Digital annotations: a formal model and its applications. In: Agosti, M. (ed.) Information Access through Search Engines and Digital Libraries, pp. 113–146. Springer, Heidelberg (in print, 2008)