Annotations and Digital Libraries: Designing Adequate Test-Beds

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. The increasing number of users and the diffusion of *Digital Libraries (DLs)* has increased the demand for newer and improved systems to give better assistance to the user during the search of resources in collections managed by *Digital Library Systems (DLSs)*. In this perspective, the annotations made on documents offer an interesting possibility for improving both the user experience of the DLS and the retrieval performance of the system itself. However, while different approaches based on annotations have been proposed, they still lack a full experimental evaluation, mainly because an experimental collection with annotation is missing. Therefore, this paper addresses the problem of setting an adequate experimental test-bed for DL search algorithms which exploit annotations, and discusses a flexible strategy for creating test collections with annotated documents.

1 Introduction

When users search a *Digital Library (DL)* they usually need answers to their information needs. They interact with the *Digital Library System (DLS)* to materialize, to the best of their abilities, their need for information. After this step, the system, interacting with the DL content, tries to retrieve the maximum possible number of documents relevant to user queries. End-users hope that the DLS can help them meet their needs by providing the documents that they are searching for. This, however, turns out to be a hard and twofold problem, because on the one hand users find it difficult to correctly explain their needs (i.e. to materialize the query to the system) and on the other DLSs have problems finding the correct resources (i.e. retrieving documents useful to the user).

Several studies have been performed to identify newer and better algorithms which aim to improve retrieval effectiveness and better satisfy end-user information needs. In this perspective, the annotations made on documents offer an interesting possibility for improving information access performance. The additional information contained in the annotations and the hypertext which connects annotations to documents are exploited to define search strategies which

* The work of Tullio Coppotelli has been partially supported by a grant of MicroWave Network S.p.A., Padua, Italy.

D.H.-L. Goh et al. (Eds.): ICADL 2007, LNCS 4822, pp. 150–159, 2007.
© Springer-Verlag Berlin Heidelberg 2007

merge multiple sources of evidence, thus increasing system effectiveness and helping users to meet their needs. DLSs which implement these new techniques require the correct evaluation of both the user interaction with the system and system effectiveness. However, while different approaches based on annotations have been proposed, they still lack a full experimental evaluation; this is because an adequate test collection is missing. Without a test collection with annotated documents setting up a correct evaluation test-bed and comparing DLSs that use annotations with DLSs that do not is impossible. It is then hard to decide if these approaches introduce improvements and which of them work better. This paper focuses on the effectiveness aspect and addresses the problem of designing an adequate experimental test-bed to evaluate DL search functionality which exploit annotations. The next section overviews related work. Section 3 describes the main characteristics of our approach. Section 4 presents two algorithms that cooperate to create the annotated test-collection. Finally, Section 5 concludes the paper.

2 Annotations and Digital Libraries

The concept of annotation is complex and multifaceted and covers a wide range of different areas. Annotations can be considered as metadata: additional data which concern an existing content and clarify the properties and the semantics of the annotated content. In this sense, annotations have to conform to some specifications which define the structure, the semantics, the syntax, and, maybe, the values annotations can assume. On the other hand, annotations can be regarded as an additional content which concerns an existing content: they increase the existing content by providing an additional layer of elucidation and explanation.

Full advantage of annotations can be taken by providing a DLS with annotation capability [1]. The primary effect of introducing annotations is to enrich the DL content; for example, by using annotations the content of a document can be broadened with personal considerations to propose different points of view or to underline text passages that need further discussion. In addition, annotations allow users to actively integrate DLs into their way of working to create a cooperative environment where annotations become the medium for users to communicate with each other.

Another important characteristic of annotations is their heterogeneity. Annotations in DLs are created by different authors with different backgrounds and at different times: the user who annotates a document may know more recent information then the author about the topic; he or she may disagree with the document content and would like to communicate this different opinion to the readers of the document. This heterogeneity is a key-point that allows for 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 queries, a feature which is so important in document retrieval.

Finally, different media can be annotated such as text, video or images and annotations themselves can be multimedia objects. However, this study focuses on the use of textual annotations to annotate textual documents. For an extensive study on annotations and their formal definition refer to [2].

Golovchinsky et al. [4] proposed the use of highlight annotations 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 limits annotation to only one facet: their use as a relevance feedback. Frommholz et al. [5] proposed a system that implements annotations for collaboration among scholars. Annotations were used to provide advanced content and content-based access to the underlying digital repository. This work adopted a broader view on annotations and enables the creation of a collaborative experience over the DL (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 multiple DLs. Annotations were used to naturally merge and link personal contents with the information resources provided by the DLSs and were exploited during the research not only to rank documents better but also to retrieve more relevant documents. This study lacks an extended evaluation but it is the first which calls for a test collection that could enable the evaluation of effectiveness of these kinds of systems.

3 A New Approach to Test-Bed Design

The use of test collections to evaluate the effectiveness of DLSs is a commonly accepted practice. Existing evaluation campaigns (TREC[1], CLEF[2], NTCIR[3]) have provided a wide range of test collections. They are reusable, produce reproducible results, encourage collaboration among researchers and cross comparison of system performance. Although these collections are general purpose and are suitable for a wide range of systems, when it comes to the need to evaluate new techniques it is possible that an adequate test collection is still lacking and consequently needs to be created. Nevertheless, creating a new collection could be itself a hard task requiring resources and time. Therefore, it can be useful to accomplish an intermediate step that still allows a reliable evaluation of the effectiveness using an alternative technique as proposed for example in [6]. Carterette et al. [7] observed that this can be the case when a researcher is performing a preliminary investigation of a new retrieval task. Hence, when it comes to the evaluation of DLSs with annotated documents, the use of an alternative technique for collection creation is a viable option.

The usual approach to test collection creation, the TREC approach [8], requires: 1) finding and acquiring a suitable set of documents; 2) manually creating annotations and topics; and 3) evaluating the relevance of documents to each topic, i.e. deciding which documents are relevant to those topics. In the case of an annotated collection: 1) assessors cannot be used to manually create the set

[1] http://trec.nist.gov/

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

[3] http://research.nii.ac.jp/ntcir/index-en.html

of annotations because a wide range of annotations written by different authors in different periods of time is needed to maintain their heterogeneous nature; 2) the pooling method is used to reduce the number of documents which human assessors need to assess for each topic. This method relies on the existence of a certain number of experiments but, in the case of annotated documents, the lack of these experiments prevents us from applying this method; hence it would be necessary to judge the relevance of each document to each topic. All these additional issues make the creation of an annotated test collection particularly difficult and, once again, confirm the need for new strategies.

This paper deals with these problems and presents a new strategy that enables the fast creation of a reliable test collection with annotated documents. The proposed technique requires starting from an already existing test collection and then creating a parallel collection of related annotations over it. These annotations are human written documents themselves that are matched to other documents on the basis of objective features, thus trying to simulate the behaviour of a human annotator that does not simply underline some passages but annotates passages of document with extensive annotations. The only constraint for the starting collection is that the documents have to be objectively divisible in more than one set (the motivation will be clarified later in this section). This strategy is not limited to the creation of a single collection: by using as a starting point collections with different characteristics, monolingual or multilingual, general or specialized, it enables the multiple creation of new collections that inherit the characteristics of the original ones. This strategy reduces the overall effort of the collection creation and has the following advantages: 1) the results of systems evaluation with the new collection are directly comparable with the results previously obtained with the original one; the testing of the systems with both collections enables the direct comparison between systems that use annotations and systems that do not, allowing the evaluation of improvements and in general the impact that new algorithms or their refinements have in DLSs; 2) it exploits existing pools to deal with a sufficient number of experiments without the expensive need for a new assessment; 3) it allows the fast creation of multiple collections with different characteristics and the evaluation of the algorithms in different contexts; and 4) it respects the heterogeneous nature of annotations.

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 [9]. The annotated collection is $C' = (D'_1, T, J)$, where D'_1 contains exactly the same document 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 strategy 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 1, 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 represents a human assessment i.e. a path between annotation $\hat{a}$ and document d passing through topic t means that a person assessed both $\hat{a}$ and d relevant to t. 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.

4 The Two Cooperating Algorithms

In this Section we introduce the two cooperating algorithms; in 4.1, we shortly outline a method that uses the human information contained in the original collection to introduce annotation in C, because full details where given in [10]. In 4.2, a new automatic technique is proposed which can be partnered to the one of 4.1 to discover an additional set of annotations.

4.1 Exploiting Assessor Assessments

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 proposed algorithm makes use of the human relevance assessments in C for matching candidate annotations with documents. The first aim of the algorithm is to match each candidate annotation $\hat{a}$ with the most suitable document d, bringing to the surface the relationship between documents. These matches respect the annotation constraint proposed in [3], i.e. one annotation can annotate only one document, and 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. If a match is possible then $\hat{a} \in A$ otherwise $\hat{a} \in (\hat{A} - A)$ and, at this point, cannot be used as a real annotation.

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 (see Figure 1): 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

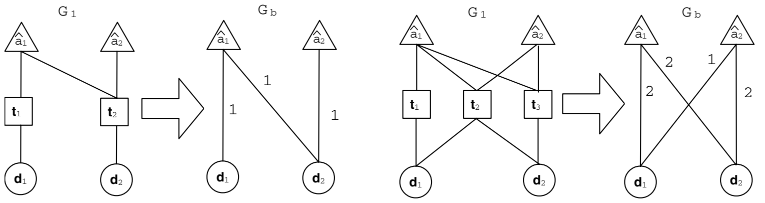

Fig. 1. Examples of the construction of graph G_b, starting from graph G_1

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. Once G_b is constructed, the second phase of the algorithm works on G_b to reach the two aims described above (a detailed presentation has been presented in [10]).

4.2 Exploiting the Whole Pool

The method suggested in 4.1 has the great advantage of bringing to the surface hidden human work, matching annotations with documents using the judgements that human assessors gave during relevance assessments. With this approach only the candidate annotations that have been judged relevant to at least one topic can be matched; it is important to note that it cannot find any match for all the other candidate annotations. Although in our preliminary experiments the number of couples (document, annotation) that this method could find was promising, in the original collection there were still a certain number of good couples that were not matched. Hence we propose an automatic technique that tries to bring to light these couples.

So far only the information about the relevance of documents to topics has been used and other information contained in C was discarded. Now we focus on the pool of documents and particularly on the reasons that caused the documents to be inserted in the pool for a certain topic (and at a later time assessed). Since we utilized one of the CLEF test collections for our initial experiments we are in the condition to exemplify by referring to that collection. The following topic is useful to illustrate how information previously discarded now can be used: "Alberto Tomba's skiing victories". As usually happens with the pooling method, the pool corresponding to that topic contains both relevant and not relevant documents; it includes not only documents about Alberto Tomba's skiing victories but also documents about skiing competitions where Alberto Tomba did not participate, those where he participated without winning and documents about his social life. While these documents have been judged not relevant to the topic and then are useless for the previous algorithm, they still contain useful information that can be used to introduce new annotations in C' coupling documents and annotations about the same subject matter, e.g. the social life of Alberto Tomba. With this aim, a complementary algorithm is proposed that first creates more paths between candidate annotations and documents building the graph using the whole pool and then brings to light the most interesting.

We define E as the set of all edges that can be created using the whole pool: for each topic t an edge e is added to the graph between t and each document d_i or annotation a_j such that $\forall i, j$ d_i and a_j are in the pool of topic t (the previous set E_1 is a subset of E). A_2 is the set of actual annotations matched with the previous algorithm and E_2 is the subset of E incident with A_2. A new graph $G_2 = (V - A_2, E - E_2)$ is then obtained using the whole pool and removing, due to the annotation constraint, all the candidate annotations already matched by the previous algorithm. Starting from G_2 a bipartite graph G_{b2} is built using the topics as connections. The main difference with the previous method is that these edges no longer reflect human assessments. The drawback of the choice to include all the documents of the pool is that there are paths in the graph that are not suitable for use as annotations for any documents. As a consequence good annotations with the previous algorithm can no longer be identified and a new strategy is required to evaluate the quality of the relationships between candidate annotations and documents and to decide which edges can be used to annotate a document. With this goal in mind, four evaluation parameters are introduced and their score is merged to compute a unique weight for each edge of G_{b2}: the affinity between topics, the score obtained using an *Information Retrieval Tool (IRT)*, the annotation generality and their temporal nearness. Each parameter measures a different aspect of the relationship between documents and candidate annotations, and their union permits an objective measure of annotation suitability. This algorithm can no longer match all annotations with document, but it does aim to annotate the greatest possible number of documents with good quality annotations; the very poor quality of some candidate annotations prevents their use as annotations, even if some of them could annotate non-annotated documents.

The affinity between documents and candidate annotations is a score P_a, ranging, like the other parameters, between 0 and 1. It uses G_{b2} structure to measure the superimposition in the content of two or more topics and the similarity of documents involved. The probability that two documents cover the same subject matter increases when the affinity increases, while edges incident with vertices with very low affinity lead to bad annotations. The formal definition of affinity between topics T_i and T_j $(i \neq j)$ is:

$$P_a^{(ij)} = \frac{|T_{ij}|}{max(|T_{ij}|)}\texttt{, where } T_{ij} = \{(\hat{a}, d) \subseteq \hat{A} \times D_1 | \hat{a} \in \hat{A}_i \cap \hat{A}_j \texttt{, and } d \in D_i \cap D_j\}$$

where $D_i, D_j \subseteq D_1$ are the sets of documents that are in the pool for these topics and $\hat{A}_i, \hat{A}_j \subseteq \hat{A}$ are the sets of annotations in the same pool.

The pooling method used to create the pool of documents in the original collection, and then to create the graph G_{b2}, selects the documents that enter the pool using the experiments run by different systems. These experiments contain information about both the ranking of the documents and their score. Although this information is discarded after the creation of the pool, it could be used to weight each edge in the graph G_{b2}. It is then useful to recompute these scores using an IRT, focusing directly on the relevance of a document to an annotation without considering their relevance to a topic. The idea is to bring to the surface documents that cover the same subject matter but that are not relevant to the same topic. High scores along not relevant paths do not contradict the assessor assessments because assessors only judged the relevance to a given topic while documents can still be related to some other topic not considered in the original collection, like in the previous VIP example. This score is computed by first creating an index over all the candidate annotations and then querying the system using the content of each document as query. In this way we obtain an ordered list of possible annotations for each document. The first K annotations of that list are considered valid while the edges between documents and annotations not in the list are deleted from G_{b2} because their relationship is too weak, i.e. the superimposition on their content is low. The use of an IRT obtained the twofold result of eliminating from G_{b2} weak edges and weighting those remaining with the score P_{ir} assigned by the IRT.

The generality score P_g is computed based on the inverse number of edges incident to the annotation's vertex, that is, the number of topics per annotation. In G_{b2} it is no longer true that increasing the number of topics in which a couple (document, annotation) belongs also increases the quality of that couple; it is only the generality of the annotation which increases. An annotation included in a lot of topic pools necessarily has to be a generic one.

The last parameter, P_t, measures the temporal nearness between documents and annotations, regardless of their order (it does not matter which comes first). The probability that documents and annotations cover the same subject matter increases when the temporal nearness increases. It is more probable to find good matches considering documents and candidate annotations temporally near.

Once these parameters have been defined, it is convenient to compute a unique score to evaluate the strength of each (document, annotation) couple: Score $S = \alpha_a * P_a + \alpha_{ir} * P_{ir} + \alpha_g * P_g + \alpha_t * P_t$ with $\alpha_a + \alpha_{ir} + \alpha_g + \alpha_t = 1$. The discriminating power of these parameters is not equal so α_a, α_{ir}, α_g and α_t have been introduced to correctly weight their importance. These weights have to be set depending on the original test collection, but it seems convenient to use a high weight for the IRT score (the most important) and a lower weight for the generality score (the less important one). Once a unique weight S is computed the maximum number of couples of vertices from the graph G_{b2} needs to be selected taking into account their quality. The algorithm presented in Section 4.1 cannot be applied because now the matching problem is more complex. The main difficulty is that there is a trade-off between the number of documents that can be annotated and the quality of these annotations: if the algorithm simply selects all possible annotations ignoring their weight, the result would be a collection with annotations of poor quality while, on the other hand, selecting only the

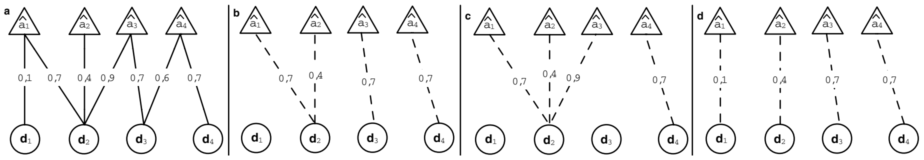

Fig. 2. From left: a is the starting graph, b is obtained with the suggested algorithm, c is obtained maximizing the score and d maximizing the documents coverage

best matches, very few documents could be annotated, reducing the advantage of this new approach.

A new greedy algorithm is proposed that resolves this trade off by proceeding in phases and trying to maximize both the number and the quality of the annotations. The examples in Figure 2 help to understand the goals of the algorithm. Starting from the input graph in Figure 2a it is possible to maximize the score of the edges (Figure 2c), obtaining a total score of 2.7 with the drawback of annotating only 2 documents over 4. The choice to maximize the number of annotated documents (Figure 2d) instead obtains a score of 1.9 with 4 documents annotated. The proposed strategy deals with the trade-off by obtaining the comprehensive score of 2.5, annotating 3 documents. The result is the annotation of an average number of documents over the two presented choices, paying only a negligible loss in the overall quality of the selected annotations. To obtain this result, the annotations that can annotate only one document are selected first. Because these annotations cannot annotate other documents, only in this case is it sufficient to apply a threshold to their quality, selecting all the edges over this threshold. In a second phase the annotations of good quality are preferred. In this phase it is important not to make counterproductive choices and only those couples $(\hat{a}_i, d_j)$ are accepted where the document d_j could not be annotated by other good annotations. The third phase is the most complex. In this phase the algorithm selects only those edges that allow it to annotate documents that are not already annotated. For each d_i not annotated a search is made for the best annotation $\hat{a}_j$ and then for the best document d_k that this annotation could annotate. If $i = k$ then the following statements are true: d_i is not already annotated, a couple $(\hat{a}_n, d_i)$ or $(\hat{a}_j, d_n)$ does not exist with a score better than that of $(\hat{a}_j, d_i)$. When these statements are verified the couple $(\hat{a}_j, d_i)$ is the best choice and if the weight of that arc is over a threshold, the document can be annotated. After these phases, to annotate more documents the algorithm relaxes the constraint $i = k$, allowing the selection of those couples where $|P(\hat{a}_j, d_k) - P(\hat{a}_j, d_i)| < \epsilon$, defining $P(\hat{a}_j, d_k)$ and $P(\hat{a}_j, d_i)$ as the weights of edges $(\hat{a}_j, d_k)$ and $(\hat{a}_j, d_i)$. With this relaxed constraint other good annotations can be introduced into the collection, thus permitting a small increase in the probability of making bad choices, that is, moving away from the optimum solution. This phase is iterated increasing the value of ϵ although at a certain point it would no longer be possible to annotate new documents. The last phase then relaxes the other constraint allowing the annotation of already annotated documents. In this phase another threshold is used to avoid the presentation of bad annotations to the assessor. By correctly choosing the parameters and thresholds of this algorithm the best fitting mixture of good quality annotation and annotated documents can be obtained, and even if the solution is not the optimum one, it is adequate for the problem we need to solve. Once the algorithm has produced a set of annotations, a human assessor has the possibility of evaluating these annotations to ensure collection reliability.

5 Conclusions

An approach to automatically create a test collection with annotated documents has been proposed, using two innovative algorithms. The preliminary results and the manual inspection of the created annotation test collection have confirmed its quality. Future work intends to complete the evaluation of the proposed approach also taking into account some of the comments of the reviewers, which need further investigation to be fully answered.

References

1. Agosti, M., Ferro, N.: Annotations: Enriching a Digital Library. In: Koch, T., Sølvberg, I.T. (eds.) ECDL 2003. LNCS, vol. 2769, pp. 88–100. Springer, Heidelberg (2003)
2. Agosti, M., Ferro, N.: A Formal Model of Annotations of Digital Content. ACM Trans. on Information Systems (TOIS) 26(1), 1–55 (2008)
3. 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)
4. Golovchinsky, G., Price, M.N., Schilit, B.N.: From Reading to Retrieval: Freeform Ink Annotations as Queries. In: Proc. 22nd Annual Int. ACM SIGIR Conf. on Research and Development in IR, pp. 19–25. ACM Press, New York (1999)
5. Frommholz, I., Brocks, H., Thiel, U., Neuhold, E., Iannone, L., Semeraro, G., Berardi, M., Ceci, M.: Document-Centered Collaboration for Scholars in the Humanities–The COLLATE System. In: Koch, T., Sølvberg, I.T. (eds.) ECDL 2003. LNCS, vol. 2769, pp. 434–445. Springer, Heidelberg (2003)
6. Sanderson, M., Joho, H.: Forming test collections with no system pooling. In: SIGIR 2004: Proc. of the 27th Annual Int. ACM SIGIR Conf. on Research and Development in IR, pp. 33–40. ACM Press, New York (2004)
7. Carterette, B., Allan, J., Sitaraman, R.: Minimal test collections for retrieval evaluation. In: SIGIR 2006: Proc. of the 29th Annual Int. ACM SIGIR Conf. on Research and development in IR, pp. 268–275. ACM Press, New York (2006)
8. Voorhees, E.M., Harman, D.K.: TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing). MIT Press, Cambridge (2005)
9. 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, Dept of Information Engineering, Univ. Padua (2006)
10. Agosti, M., Coppotelli, T., Ferro, N., Pretto, L.: Exploiting Relevance Assessment for the Creation of an Experimental Test Collection to Evaluate Systems that Use Annotations. In: DELOS Conference 2007, ISTI-CNR, Pisa, Italy, pp. 195–202 (2007)