A Hybrid Approach for Multi-Faceted IR in Multimodal Domain

Serwah Sabetghadam, Ralf Bierig, Andreas Rauber

Institute of Software Technology and Interactive Systems
Vienna University of Technology
Vienna, Austria

Abstract. We present a model for multimodal information retrieval, leveraging different information sources to improve the effectiveness of a retrieval system. This method takes into account multifaceted IR in addition to the semantic relations present in data objects, which can be used to answer complex queries, combining similarity and semantic search. By providing a graph data structure and utilizing hybrid search in addition to structured search techniques, we take advantage of relations in data to improve retrieval. We tested the model with ImageCLEF 2011 Wikipedia collection, as a multimodal benchmark data collection, for an image retrieval task.

Keywords: Multimodal, Information Retrieval, Graph, Hybrid Search, Facet, Spreading Activation

1 Introduction

The web is increasingly turning into a multimodal content delivery platform. This trend creates severe challenges for information retrieval. Using different modalities —text, image, audio or video—to improve an IR System is challenging since each modality has a different concept of similarity underneath.

There are numerous related works in this area, e.g., in combination of text and images, given the massive web data, relevant web images can be readily obtained by using keyword based search [7, 5]. Utilizing intermodal analysis for automatic document annotation [11] is another possibility.

In addition to the observation that data consumption today is highly multimodal, it is also clear that data is now heavily semantically interlinked. This can be through social networks (text, images, videos of users on LinkedIn, Facebook or the like), or through the nature of the data itself (e.g. patent documents connected by their metadata - inventors, companies). Connected data poses structured IR as an option for retrieving more relevant data objects.

We observe, since 2005, a trend towards hybrid search, leveraging both structured and un-structured IR [8, 4, 6]. Combining the two search methods is challenging because of their respective diversity. In unstructured IR we have multimodality – the diverse nature of the data objects, while in structured IR we have multi-connectivity – the diverse nature of the links of the graph.

In this paper, we propose a model, named Astera, to leverage hybrid search in order to handle the diverse nature of the nodes and edges in the multimodal content domain. We model domain specific collections with the help of different relation types, and enrich the available data by extracting inherent information in the form of facets. Our model is a triangle of hybrid search, faceted search and multimodal data.

We show the applicability of this model on the multimodal domain by using the ImageCLEF 2011 Wikipedia collection dataset [17]. We perform a basic yet thorough evaluation and show that our model matches the efficiency of non-graph based indexes, while having the potential to exploit different facets for better retrieval. We show that the result of multimodal faceted approach, excels baseline results.

The paper is structured as follows: in the next section, we address the related work, followed in Section 3 by the basic definition of our model, graph traversal and weighting. The experiment design is shown in Section 4. The results are discussed in Section 5, and finally, conclusions and future work are presented in Section 6.

2 Related Work

Astera is at the crossroad of different related work areas: multimodal retrieval, hybrid search, faceted and semantic search. We try in this section to clarify the differentiation of Astera towards each category and highlight its new message.

There are many efforts in multimodal retrieval, e.g. in combining textual and visual modalities. Martinent et al. [11] propose to generate automatic document annotations from inter-modal analysis. They consider visual feature vectors and annotation keywords as binary random variables. Srinivasan and Slaney [16] add content based information to image characteristics as visual information to improve their performance. Their model is based on random walks on bipartite graphs of joint model of images and textual content. I-Search, as a multimodal search engine [9], defines relations between different modalities of an information object, e.g. a lion's image, its sound and its 3D representation. They define neighbourhood relation between two multimodal objects which are similar in at least one of their modalities. However, in I-Search, the semantic relation between objects (e.g. a dog and a cat object) is not considered.

In combining structured and unstructured IR, Magatti [10] provides a combination of graph and content search. For example, in an organization, members have hierarchal relations by their roles, meanwhile there are related documents to them. The structured search engine NAGA [8], provides the results of a structured (not keyword) query by using subgraph pattern on an Entity-Relationship graph. Rocha et al. [12] use spreading activation for relevance propagation applied to a semantic model of a given domain. Targeting RDF data, SIREn [4] supports both keywords and structured queries. Elbassuoni and Blanco [6] select subgraphs to match the query and do the ranking by means of statistical language models. We build upon these works and complement them with the concept of faceted search.

We extend the common notion of faceted search in order to enable a more flexible information access model. We connect extracted facets to their information objects and treat them as individual nodes. This provides various possibilities for both early and late fusion.

Another aspect of Astera is that the data model is a graph which relates it to work done in the semantic web domain. Search in the semantic web is keyword-based. Some research is particularly concerned with generating adequate interpretations of user queries [15]. In addition to semantic search, Astera is able to consider similarity computations (between object facets) for searching an information object. Furthermore, we generalize the query and provide a list of highly related neighbours for a user, rather than simply providing an exact response.

Related research on the ImageCLEF 2011 Wikipedia collection is generally based on a combination of text and image retrieval [17]. To our best knowledge, there is no approach that has modelled the collection as a graph structure and no approach has therefore leveraged the explicit links between objects and between objects and their features.

3 Model Representation

We define a model to represent information objects and their relationships, together with a general framework for computing similarity. We see the information objects as a graph $G = (V, E)$, in which V is the set of vertices (including data objects and their facets) and E is the set of edges. By facet we mean inherent information of an object, otherwise referred to as a representation of the object. For instance, an image object may have several facets (e.g. color histogram, texture representation). Each of these is a node linked to the original image object. Each object in this graph may have a number of facets. We define four types of relations between the objects in the graph. The relations and their characteristics are discussed in detail in [13]. We formally define the relation types and their weights as follows:

- **Semantic** (α): any semantic relation between two objects in the collection (e.g. the link between lyrics and a music file). The edge weight w_{uv} is made inversely proportional to the α-out-degree of the source node u (the number of outgoing α links from u). Thus $w_{uv} = 1/N_u^{(\alpha)}$. This reduces the effect of very connected nodes on the spreading process and simulates fanout constraint[3] to decrease the distributed energy to very low for popular nodes.

- **Part-of** (β): a specific type of semantic relation, indicating an object as part of another object, e.g. an image in a document. This is a containment relation as an object is part of another one, and therefore we set the default weight to 1.

- **Similarity** (γ): relation between objects with the same modality. This relation is defined just between the facets of the same type of two information objects, and the weight is the similarity value between the facets according to some facet-specific metric. For instance, we can compute the similarity between Edge Histogram facet of two images.
- **Facet** (δ): linking an object to its representation(s). In our graph traversal, we can reach an object from its facet and go to other objects but we do not walk from an object to its facets. The edge in the direction of the object to the facet is weighted 0. On the other direction, from facet to the object, weights are given by perceived information content of features, with respect to the query type. For instance, with a query like "blue flowers", the color histogram is a determining facet that should be weighted higher. These weights should be learned for a specific domain, and even for a specific query if we were to consider relevance feedback.

In addition to the edge weights just defined, we consider the use of a self-transitivity value (st) to emphasize remaining on a specific state. This value leaves part or all of the initial energy with the current node.

$$W_{v|u} = \begin{cases} (1 - st)w_{uv} & u \neq v \\ st & u = v \end{cases} \tag{1}$$

where $W_{v|u}$ is the weight of going from node u to v.

3.1 Traversal method - Spreading Activation

For traversing the graph and finding the relevant result for a query, we propose to use spreading activation (SA). The SA procedure, always starts with an initial set of activated nodes, usually the result of a first stage processing of the query. During propagation, surrounding nodes are activated and ultimately, a set of nodes with respective activation are obtained. After t steps, we use the method provided by Berthold et al. [2], to compute the nodes' activation value :

$$a^{(t)} = a^{(0)} \cdot W^t \tag{2}$$

where $a^{(0)}$ is the initial activation vector, W is the weight matrix—containing different edge type weights—, and $a^{(t)}$ is the final nodes' activation value used for ranking.

3.2 Hybrid Search

The use of results from independent modality indexing neglect a) that data objects are interlinked through different relations and b) that many relevant images can be retrieved from a given node by following semantic or 'part-of' relations. Our hybrid ranking method consists of two steps: 1) In the first step, we perform an initial search with Lucene and/or Lire to obtain a set of activation nodes. 2) In the second step, using the initial result set of data objects (with normalized scores) as seeds, we exploit the graph structure and traverse it.

We follow the weighted edges from the initiating points for t steps. We perform the spreading activation and at the end recompute the ranked result based on the activation value nodes received via propagation (Equation 2).

The number of transitions is determined by imposing different stop rules: distance constraint [3], fan-out constraint [3] or type constraint[12]. In this version of our model, we use the distance constraint to stop the traversal.

4 Experiment Design

In this section, we describe the dataset and different retrieval methods. We used ImageCLEF 2011 Wikipedia collection to evaluate the indexing of multi-modal multimedia content and to test the functionality and performance of our hybrid search method.

4.1 Data Collection

We applied the ImageCLEF 2011 Wikipedia collection as a benchmark. This collection is based on Wikipedia pages and their associated images. It is a multimodal collection and an appropriate choice for testing the rich and diverse set of relations in our model. The goal in the setting of this particular test collection is to retrieve images. Each image has one metadata file that provides information about name, location, one or more associated parent documents in up to three languages (English, German and French), and textual image annotations (i.e. caption, description and comment). The collection consists of 125,828 documents and 237,434 images. We parsed the image metadata and created nodes for all parent documents, images and corresponding facets. We created different relation types: the β relation between parent documents and images (as part of the document), and δ relation between information objects and their facets. We use the 50 English query topics.

4.2 Standard Text and Image Search

In the indexed search approach, as first phase of our hybrid search, we use Lucene indexing results both for documents and images. The computed scores in both modalities are normalized per topic between (0,1). Different indexings based on different facets are:

- **Document tf.idf facet**: We utilize default Lucene indexer, based on tf.idf, as document facet. We refer the result set of this facet as R1.
- **CEDD facet**: For image facets, we selected the Color and Edge Directivity Descriptor (CEDD) feature since it is considered the best method to extract purely visual results [1]. We refer to the image results of the CEDD facet as R2.

- **Image textual annotation tf.idf facet**: We use metadata information of the images (provided by the collection), as image textual facets (Tags). Metadata XML files of ImageCLEf 2011 Wikipedia collection, includes textual information (caption, comment and description) of images. Using Lucene we can index them as separate fields, and search based on a multi-field indexing. Tags search result make R3 result set.

Weighting Strategy Each information object (e.g. image, document or any other type of information object) may have many facets. They can receive at maximum, the score of 1 from facets. We weight visual facets with 0.3 and textual facets with 0.7 as an experimental parametrization based on a set of previous empirical tests [14].

The formula for the combined scoring is like: $obj\_score = \sum_{i=0}^{n} w_i.f_i$ where $\sum_{i=0}^{n} w_i = 1$. Variable n is the number of the facets, and w_i is the weight of facet f_i.

For images, we have visual facet of CEDD, and metadata information as textual facet. Mapped to the score formula, it is $(0.7 * Tags + 0.3 * CEDD)$. For document objects, we have textual facet tf.idf and we give $(1.0 * tf.idf)$ as weighting. The weights are fixed based on the experiments, but should be learned.

4.3 Graph Search

In this section we describe how we manage facet fusion and graph traversal.

Subgraph Traversal The ImageCLEF 2011 Wikipedia collection contains the total size of 363,262 information objects (images and documents without considering the facet nodes). With matrix in this size, we need about 983GB RAM to perform matrix multiplication. In order to make the calculation feasible for large collections, our strategy in Astera is to only contain the set of nodes that will be potentially reachable after N steps, and generate a smaller adjacency matrix only for them. However, this set of reachable nodes depends on the query . Therefore, for different query topic and different number of steps, we work with different subgraphs of the whole graph.

Starting from top ranked nodes for a query topic, we visit next round neighbours in each step. After visiting all neighbours to the specific step in the graph, we create the adjacency matrix W out of that. The cell values of the adjacency matrix are the edge weights between different visited nodes.

As shown in Equation 2, we compute the steps in the graph by matrix multiplication. The $a^{(0)}$ vector is composed of top ranked nodes of R1,R2 and R3 (as non-zero elements), and visited neighbours through traversal (as zero elements). The final vector, a^t, provides the final activation value of all nodes. We filter out the images and calculate precision and recall based on their scores. We chose 9 steps to show spreading activation behaviour in primary steps in Astera. We are visiting on average about 15,000 nodes per topic.

Maximum nodes searchable from text With adding one facet for images and documents, we add about 363,262 nodes to the collection which results in 726,524 node graph. Starting from text indexing results, we continued the traversal in the graph up to visiting no new node. This happened in average after 40 steps and visiting about 170,000 nodes of total 726,524 nodes. This shows that starting just from documents provides limited view to the collection and we miss related objects in the other parts of the graph.

Facet Fusion In practice, we are making a form of late facet fusion by combination of different scores and giving one score to the parent information object. However, it is not in the traditional way of late fusion. Since we are not making the result rank list out of top ranked nodes. We initiate their scores in graph nodes and then start propagation. In Astera, facet fusion is implicitly calculated by matrix multiplication and final vector computation.

5 Results and Discussion

5.1 Experiment 1: Baseline

The evaluation of this experiment represents our baseline and applies a standard Lucene index in combination with a standard Lucene search. For each ranked document result, we extracted its associated images and ranked them based on the score of the document. The result is shown in the first row of Table 1 (e.g. 0.311 for p@10). We additionally refine the baseline by computing the similarity between each of the query images and each of the result list images and keep the value of the maximum similarity SV as reference as shown in the following formula:

$SV_{q_{imgs},res_{img}} = max(Sim(q_{img_i}, res_{img})), 1 \leq i \leq 5$

Now each image result has two scores, the text scores and the image similarity score. By applying a range of different weightings for their linear combination, we discovered the best result is obtained by weighting text with 0.7 and images with 0.3 (see second row of Table 1)[14]. Results purely obtained from image-only searches had very low recall and are not presented here.

Table 1: Results for baseline

txt weight	*img weight*	*p@10*	*r@10*	*p@20*	*r@20*
1	0	0.311	0.105	0.247	0.129
0.7	0.3	0.345	0.109	0.281	0.133

5.2 Experiment 2: Graph Modelled Data

Having modelled the collection in a graph, we designed several experiments based on tf.idf, CEDD and Tags facets, and *st* values. We aim to examine the effect of adding image facets and combination of document and image facets with different *st* values in our graph search.

Search with document facet (R1) In this experiment we use tf.idf facet results as initiating points in the graph. We do not include any visual or textual facet of the images. From Table 2, we observe that we are receiving better precision by using the graph structured data. We are receiving about the 0.34 of baseline result for P@10.

As the activation is propagated further up to 9 steps, we observe a decrease in precision. We are receiving almost the same precision in even steps compared to their prior odd steps. The reason is that *st* holds the value of 0.9, and we count all images visited up to current state in the calculation.

Table 2: Result for documents without image facets, *st*:0.9

steps	*st*	*p@10*	*r@10*	*p@20*	*r@20*
1	0.9	**0.34**	0.136	0.25	0.161
2	0.9	0.34	0.136	0.25	0.161
3	0.9	0.286	0.114	0.208	0.158
4	0.9	0.28	0.112	0.206	0.149
5	0.9	0.252	0.104	0.188	0.144
6	0.9	0.244	0.104	0.18	0.138
7	0.9	0.218	0.095	0.176	0.138
8	0.9	0.194	0.081	0.158	0.124
9	0.9	0.19	0.08	0.148	0.115

Bipartite Graph We observe that, the collection modelled is a bipartite graph combined of images on one side and documents on the other side. There is no relation between images or between documents. Therefore, without self-transitivity (*st*) value, energy flows totally from one side to the other. Facets are not included in this interpretation, since there is the way just from facet to the images and the way back is blocked by weight 0 on the edge from the information object to its facet.

Search with CEDD facet added (R1 and R2) In this experiment, top images, based on CEDD similarity are added to the $a^{(0)}$ vector to activate the graph. The activation vector is therefore a combination of indexed documents and images results. We first consider no *st* value. In comparison to R1 result, we are receiving the worst results specially in even steps (Figure 2). The reason is that starting from top image nodes, we are visiting more images in even steps and they are mostly non relevant.

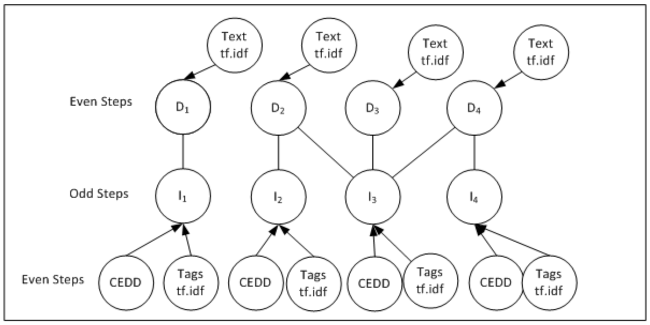

Fig. 1: Graph model: Starting from documents, we visit images in odd steps. Each image may have different facets.

Self-Transitivity added In order to increase inertia, and include image results in all steps, from here we give *st* value to all nodes. The same iterations with *st* values 0.1 and 0.9 are shown in Figure 2. This time, we see high decrease in precision, especially for value 0.9. With high *st* value, the CEDD results have a high impact in the selection of the top images. This shows that our model follows the proved claim in literature that pure image results are poorer compared with text-based results and should receive less weight.

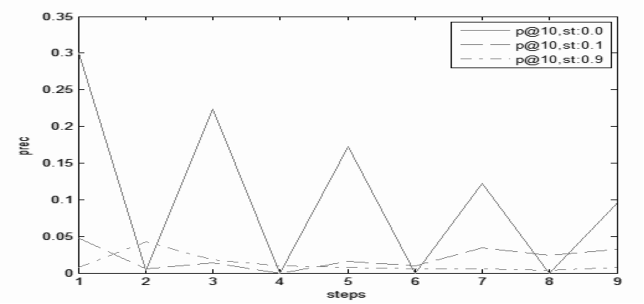

Fig. 2: Prec@10 for documents and images with CEDD facet

Weighted document and CEDD facets (R1 and R2) In order to remove the high influence of top image results in the propagation, we weight the document and image score results. In order to compare with the best result we had in baseline search (with 0.7 weight to the documents and 0.3 to the images), we perform the same here. In Table 3 we observe that the weighted result is much better in even steps than Figure 2 with *st*=0.9, because the scores of the images are reduced to match their perceived importance for retrieval. We observe almost the same precision with R1 result experiment, however, with better recall in first four steps. Going further in the graph we see more number of images which decrease the efficiency of the system.

Table 3: Result for documents and images CEDD facet, *st*:0.9

steps	*st*	*p@10*	*r@10*	*p@20*	*r@20*
1	0.9	0.344	0.135	0.25	**0.188**
2	0.9	0.338	0.133	0.257	**0.193**
3	0.9	0.29	0.115	0.207	**0.163**
4	0.9	0.266	0.101	0.175	**0.131**
5	0.9	0.234	0.093	0.165	0.122
6	0.9	0.136	0.052	0.098	0.068
7	0.9	0.11	0.039	0.077	0.055
8	0.9	0.094	0.032	0.065	0.042
9	0.9	0.084	0.056	0.062	0.038

Search with document and metadata facets (R1 and R3) In this experiment we search based on document and metadata facet results. We see an increase of 0.06 in the first step (Table 4). Also in third step we have better precision, which shows that we visit related documents, not only after one step, but also after three steps. We see that using metadata facet we have better recall in the first three steps as well. In these experiments the *st* value is 0.9 which means energy is partially remained in the nodes as well. Therefore, all image nodes visited in the traversal, participate in our calculations. Without *st* value, no energy is remained in the nodes previously visited.

Table 4: Result for documents and image metadata facet, *st*:0.9

steps	*st*	*p@10*	*r@10*	*p@20*	*r@20*
1	0.9	**0.362**	0.139	0.265	**0.189**
2	0.9	**0.346**	0.132	0.259	**0.175**
3	0.9	**0.308**	0.119	0.224	**0.165**
4	0.9	0.24	0.088	0.187	0.135
5	0.9	0.212	0.081	0.164	0.118
6	0.9	0.158	0.06	0.133	0.097
7	0.9	0.164	0.06	0.128	0.091
8	0.9	0.144	0.56	0.113	0.085
9	0.9	0.084	0.027	0.062	0.038

Search with document, CEDD and metadata facets (R1, R2 and R3) We included all three result sets in this experiment. Receiving higher recall than previous experiment with R1 and R2 shows that the combination of R2 and R3 hit points helps visiting more related nodes. Precision in the first step is the same as combination of R1 and R3 result. This means that CEDD top ranked nodes did not help. In the second step (according to Figure 1), starting from document hit nodes (R1) we visit documents again which do not affect the result of this stage. However starting from images (R2 and R3), we visit new images in second step. Precision increase to 0.372, demonstrates visiting related documents from R2 and R3 points in first step, that in second step lead to more related images.

We observe that CEDD could have positive effect in combination with Tags to increase the precision in second step (Table 5), while in the combination of R1 and R2 experiment, it was not effective (Table 3).

Table 5: Result for documents and images CEDD and metadata facets, *st*:0.9

steps	*st*	*p@10*	*r@10*	*p@20*	*r@20* ($)
1	0.9	**0.358**	0.14	0.27	**0.195**
2	0.9	**0.372**	0.137	0.272	**0.193**
3	0.9	**0.308**	0.12	0.22	**0.166**
4	0.9	0.25	0.093	0.186	0.127
5	0.9	0.218	0.083	0.162	0.113
6	0.9	0.114	0.037	0.088	0.06
7	0.9	0.114	0.037	0.085	0.056
8	0.9	0.138	0.056	0.113	0.085
9	0.9	0.068	0.055	0.107	0.083

6 Conclusion

We presented a multifaceted model for hybrid search in a multimodal domain. In this model, data collections can be described based on different link types. We enriched the modeled connections by extracting inherent information of data objects as facets. The preliminary results of combination of text and image facets show the correct functionality in the combined modalities. However, we were able to improve these results by using a weighted combination of document and image results. Furthermore, the Astera model enabled us to search the collection from different points of view by using different facets. Utilizing image textual facet increased precision and recall. Further, combination of two different facets of images (CEDD and Tags) gave better result than the sum of their individual results. This demonstrates the positive effect of the combination of different facets in Astera.

Our future work will focus on the following: 1) Learning the weight of different facets through supervised learning methods. 2) Further exploring the semantic relations between the ImageCLEF 2011 Wikipedia collection and DBPedia. For example, traversing the graph starting from the collection and spreading through DBPedia until returning to the collection, considering the effect of semantic links. 3) Using concept extraction to create additional, more meaningful semantic links between query topics and image textual annotations(caption, comment and description of the image)

References

1. T. Berber, A. H. Vahid, O. Ozturkmenoglu, R. G. Hamed, and A. Alpkocak. Demir at imageclefwiki 2011: Evaluating different weighting schemes in information retrieval. In *CLEF*, 2011.
2. M. R. Berthold, U. Brandes, T. Kotter, M. Mader, U. Nagel, and K. Thiel. Pure spreading activation is pointless. In *CIKM'09*, 2009.
3. F. Crestani. Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, 11, 1997.
4. R. Delbru, N. Toupikov, M. Catasta, and G. Tummarello. A node indexing scheme for web entity retrieval. In *ESWC*, 2010.
5. L. Duan, W. Li, I. W. Tsang, and D. Xu. Improving web image search by bag-based reranking. *IEEE Transactions on Image Processing*, 20(11), 2011.
6. S. Elbassuoni and R. Blanco. Keyword search over RDF graphs. CIKM, 2011.
7. R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *Proc. of Intl. Conf. on Computer Vision*, 2005.
8. G. Kasneci, F. Suchanek, G. Ifrim, M. Ramanath, and G. Weikum. Naga: Searching and ranking knowledge. In *ICDE*, 2008.
9. M. Lazaridis, A. Axenopoulos, D. Rafailidis, and P. Daras. Multimedia search and retrieval using multimodal annotation propagation and indexing techniques. *Signal Processing: Image Comm.*, 2012.
10. D. Magatti, F. Steinke, M. Bundschus, and V. Tresp. Combined Structured and Keyword-Based Search in Textually Enriched Entity-Relationship Graphs. In *Proceedings of the Workshop on Automated Knowledge Base Construction*, 2011.
11. J. Martinet and S. Satoh. An information theoretic approach for automatic document annotation from intermodal analysis. In *Workshop on Multimodal Information Retrieval*, 2007.
12. C. Rocha, D. Schwabe, and M. P. Aragao. A hybrid approach for searching in the semantic web. WWW, 2004.
13. S. Sabetghadam, M. Lupu, and A. Rauber. Astera - a generic model for multimodal information retrieval. In *Proc. of Integrating IR Technologies for Professional Search Workshop*, 2013.
14. S. Sabetghadam, M. Lupu, and A. Rauber. A combined approach of structured and non-structured IR in multimodal domain. In *ICMR*, 2014.
15. S. Shekarpour, S. Auer, A. Ngomo, D. Gerber, S. Hellmann, and C. Stadler. Keyword-driven sparql query generation leveraging background knowledge. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE/WIC/ACM International Conference on Web Intelligence*, volume 1. IEEE, 2011.
16. S. Srinivasan and M. Slaney. A bipartite graph model for associating images and text. In *Workshop on Multimodal Information Retrieval*, 2007.
17. T. Tsikrika, A. Popescu, and J. Kludas. Overview of the wikipedia image retrieval task at imageclef 2011. In *CLEF*, 2011.