A Combined Approach of Structured and Non-structured IR in Multimodal Domain

Serwah Sabetghadam, Mihai Lupu, Ralf Bierig, Andreas Rauber
Institute of Software Technology and Interactive Systems
Vienna University of Technology
{sabetghadam,lupu,bierig,rauber}@ifs.tuwien.ac.at

ABSTRACT

We present a generic model for multimodal information retrieval, leveraging different information sources to improve the effectiveness of a retrieval system. The proposed method is able to take into account both explicit and latent semantics present in the data and can be used to answer complex queries, not currently answerable neither by document retrieval systems, nor by semantic web systems. By providing a hybrid approach combining IR and structured search techniques, we prepare a framework applicable to multimodal data collections. To test its effectiveness, we instantiate the model for an image retrieval task.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: General ; H.3.3 [**Information Search and Retrieval**]: Metrics—*Retrieval models, Search process*

General Terms

Design, Experimentation

Keywords

IR, multimodal, graph, spreading activation

1. INTRODUCTION

Multimodal IR has become one of the challenges in IR domain. Getting help from different modalities—text, image, audio or video—in order to provide better results to satisfy the users' information needs is difficult because of the different concepts of similarity in each of these modalities. 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 [5, 7]. Utilizing intermodal analysis for automatic document annotation [11] is another attempt in this area. In addition to the observation that data consumption today is multimodal, it is also clear that data is now heavily interlinked. This can be through social networks (text, images, videos on LinkedIn, Facebook or the like), or through the nature of the data itself (e.g. patent documents connected by their metadata - inventors, companies). 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 problematic because of their respective diversity. In unstructured IR we have multi-modality – the diverse nature of the data objects, while in structured IR we have multi-connectivity – the diverse nature of the links of the graph. We hypothesize that the diverse nature of the nodes and edges is in fact better handled together and propose Astera, as a model for multimodal IR. Astera models domain specific collections with help of different relation types, and enriches the available data by extracting inherent[1] information of data objects.

In this paper, we further describe Astera and show initial experiments. We show the applicability of Astera on multimodal domain by using the ImageCLEF 2011 dataset. We perform a basic yet thorough test and show that Astera matches the efficiency of non-graph based indexes, while having the potential to exploit semantic relations in further experiments.

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

2. RELATED WORK

There are many efforts in combining textual and visual modalities. Martinent et al. [11] suggest to generate automatic document annotations from inter-modal analysis, considering visual feature vectors and annotation keywords as binary random variables. Srinivasan and Slaney [15] add content based information to image characteristics to improve their performance. I-Search is a multimodal search engine project [9], in which a multimodality relation is defined between different modalities of an information object, e.g. a dog image, its sound (barking) and its 3D representation. They define a neighbourhood relation between two multimodal objects which are similar in at least one of their modalities. This type of relation is modelled in Astera via similarity relation types. However, in I-Search, neither semantic relation between objects (e.g. a dog and a cat object) is considered, nor the importance of these relations in answering the user's query.

[1] here, by inherent we mean the kind of information extracted from a data object

ICMR '14 April 01 - 04 2014, Glasgow, United Kingdom Copyright 2014 ACM 978-1-4503-2782-4/14/04 ...$15.00.

From the search method point of view, a number of hybrid search systems have already been worked on. Targeting RDF data, Elbassuoni and Blanco [6] select subgraphs to match the query and rank by means of statistical language models. As a hybrid Web search framework, SIREn [4] supports both keywords and structured queries over RDF data. They try to provide more efficiency for user through novel indexing scheme. Magatti [10] provides a model based on Entity-Relationship graphs in which nodes are connected to unstructured data. He uses SPARQL query to filter the keyword search result. Tonon et.al. [16] using a hybrid search on Linked Open Data try to retrieve better result by exploring selected semantic links. In Astera we provide a hybrid search model that is not limited to work on RDF data.

3. MODEL REPRESENTATION

Astera is independent of the data modalities, as long as a similarity function may be calculated between objects of the same modality. Astera can model domain specific multimodal collections. The relations between data objects are modelled in a graph $G = (V, E)$; V is the set of vertices comprising of data objects and their facets; E is the set of edges. By Facet we understand information inherent to the 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.

The relations and their characteristics are discussed in detail in [14]. We provide the definitions here, for readability:

- Semantic: any semantic relation between two objects in the collection (e.g. the link between a lyric and a music file)
- Part-of: a specific type of semantic relation, indicating an object as part of another object, e.g. an image in a document.
- Similarity: between objects with the same modality.
- Facet: linking an object to its representation(s).

3.1 Graph Traversal

For traversing the graph and finding the relevant result for a query, we propose to use spreading activation (SA). This method is inspired by simulated neural networks, however, in SA we do not have training phase. Edge weights are defined based on the semantics of the modelled domain. The SA procedure always starts with an initial set of activated nodes. Different values can be given to the initial nodes according to the task being solved. They are usually the result of a first stage processing of the query, e.g. a distance measure between the objects and the query. During propagation other nodes get activated and ultimately, a set of nodes with respective activation is obtained.

In what follows, we denote the initial activation of the nodes as $a^{(0)}$ and the activation in t-th iteration as $a^{(t)}$. The input value in_v for each node v is the aggregation of output values of its neighbours [2]:

$$in_v^{(t)} = \sum_{u \in V} o_u^{(t-1)} \cdot W_{u,v} \qquad (1)$$

where $W_{u,v}$ is the edge weight between nodes u and v in the weight matrix W. Different functions can be used on input value to activate the node, like linear, sigmoid or step function [3].

$$a_v^{(t)} = act(in_v^{(t)}) \qquad (2)$$

In next step to compute output of a node, an output function can be applied on the activation function result.

$$o_v^{(t)} = out(a_v^{(t)}) \qquad (3)$$

Based on Equation 3, the output of a node in SA is the result of applying the activation and output functions on the input value of the node. If the input function is defined as linear combination and the output and activation functions are identity functions, then the Equation 3 in SA can be written as $a_v^{(t+1)} = \sum_{u \in V} a_u^{(t)} \cdot W_{uv}$, which in compact form is:

$$a^{(t+1)} = a^{(0)} \cdot W^{t+1} \qquad (4)$$

The important part is how we define the weight matrix W. For each type of edge, we have an independent definition:

Semantic: Our method for this kind of relation follows Rocha's work [13]. The weight is defined based on the number of semantic relations between two nodes. $w_{jk} = N_{jik}/N_{ij}$, where N_{jik} represents the number of objects i that both nodes of j and k are related to, and N_{ij} is the number of objects related to object j.

Part-of: Since in this relation an object is part of another object, then the weight is given as 1.

Similarity: This relation is defined just between the facets of two objects from the same type.

Facet: The edge in the direction of the object to the facet is weighted 0 and on the other direction, from facet to the object, is weighted 1 because in our graph traversal we do not walk from an object to its facet, but we can reach an object from its facet and go to other objects.

3.2 Hybrid Search Method

The retrieval procedure in Astera consists of two phases. First, an initial result set R_1 is obtained from standard indices, based on different query facets. Second, starting from R_1, we generate the set of related nodes R_2. For example, if the query is the combination of text and image, then two lists of top n indexed results are obtained based on text facets and image facets, targeting different nodes in the graph. From each of these nodes, SA is started in parallel and executed for a number of t steps. Facet fusion is implicitly calculated by matrix multiplication and final vector computation.

This number of transitions is determined by imposing different stop rules: distance constraint [3], fan-out constraint [3] or type constraint[13]. In this version of Astera, we use the distance constraint to stop the traversal.

4. EXPERIMENT DESIGN

4.1 Data preparation

The benchmark data collection used is ImageCLEF 2011[2], which is based on wikipedia pages and contained images. The Wikipedia image retrieval task investigates how multimodal image retrieval approaches combine textual and visual features to satisfy user information need.

We chose this collection as our test-bed because it is multimodal and covers the diverse relation types defined in Astera. The collection contains text files, i.e. wiki pages, and the images inside them. Each image has a metadata file containing its name, file address, parent documents in three languages (en, de, fr) if available, caption, description and comment of the image in these three languages. The nodes created per image in Astera to model an image metadata is shown in Figure 1.

[2] http://www.imageclef.org/wikidata

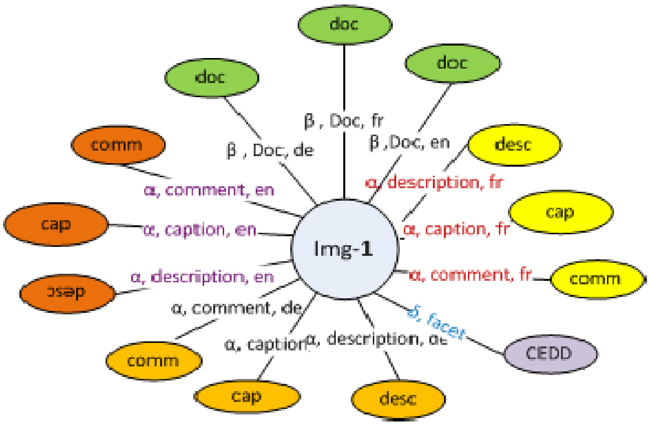

Figure 1: Image metadata extracted relations, modelled in Astera

The collection consists of 237,434 images and their related textual annotations. From four feature sets available for the images (CEDD - Color and Edge Directivity Descriptor, CIME, TLEP - Texture Local Edge Pattern, and SURF- Speeded Up Robust Features), CEDD is chosen for similarity computations in Astera because, based on the result of Berber et al. experiments [1], the best purely visual results are obtained using this descriptor.

From different languages of query topics we use the English version. For each query, Lucene provides us results and we start activating the graph based on top ranked ones. We make the matrix of the graph in Matlab. Our matrix consists of all nodes seen in the number of steps traversed in the graph. In the initial vector $a^{(0)}$ items are non zero for elements in R_1, and zero for all other elements. We evaluate the ranked list based on Equation 4.

Adding DBPedia. In order to make the ImageCLEF 2011 collection more connected and affecting potential semantic relations between collection objects, we connect the wiki collection to DBpedia through their equivalent pages. We downloaded the corresponding DBpedia dump. The ImageCLEF 2011 Wikipedia collection uses the ImageCLEF 2010 Wikipedia Collection [12], which is based on the September 2009 Wikipedia dumps. Therefore we downloaded DBpedia version 3.4 which is based on Wiki dump September 2009.

For each object in the DBpedia dataset, we create all available semantic links in Astera. Each triple in DBpedia RDF is in the form of source, predicate, literal/source. Before adding any source node we check if we had already this node in the graph to use the existing nodes as source node. By adding all DBpedia pages as source and literal nodes to our graph, a more connected large scale graph is obtained (Figure 2).

4.2 Experiment 1: Baseline Data

In this experiment, we do not use the graph model and just evaluate the Lucene results with their related images. As in content based image retrieval (CBIR), this phase consists of two steps: first is text search in which all documents are searched based on the given query (topic). The top 100 document per topic are retrieved. We find the images related to each of these documents and rank based on their document score. The result of this phase is shown in Table 1 where the text weight is 1 and image weight is 0; the precision value p@10 is 0.311.

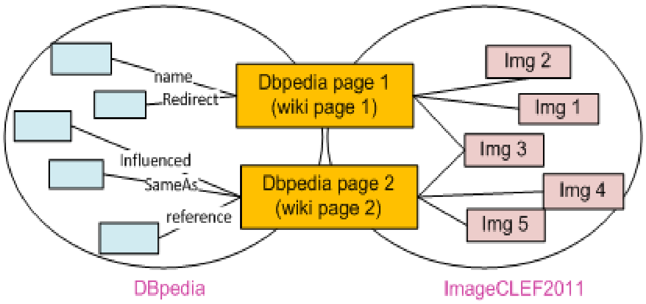

Figure 2: Linking ImageCLEF 2011 data objects to DBpedia

Table 1: Result for static data collection

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

In the next step, we compute the similarity between each of the query images and each of result list images and keep the max similarity value SV as the reference value. The similarity value is computed as:

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

Choosing max similarity is based on choosing the highest value of similarity between query topic images and result images as more relevant one. By giving different weightings to text and image for linear combination, the best result, 0.345 for p@10 is obtained by 0.7 assigned to the text weight and 0.3 assigned as image weight.

4.3 Experiment 2: Graph Structured Data

We modelled the ImageCLEF 2011 in a graph, based on the nodes in Figure 1. Next experiments are performed on this graph model.

4.3.1 Documents not scored

In this experiment, we give the same weight to all edge types of a node. Hence, all images of a document receive the same weight of $1/nb$ where nb is the number of node neighbours. We perform this weighting for all nodes seen in the number of steps we go further in the graph: for each node we observe as a new neighbour we get all its neighbours and give equal weight to correspondent edges. The only exception here is that we give weight zero to the edge from image to its facet, but treat the facet nodes the same as other nodes. Since facets are just connected to their images, they have weight one on the edge to their image. The result of graph traversal after 1, 2 and 3 steps is shown in Table 2.

4.3.2 Documents and Images, not scored

In this experiment we include image similarity results in initiating the graph search. We compute the image similarity of each topic images with all images in the collection. The top m images for each topic plus the top n documents from Lucene result contain the value of $1/(n + m)$ in the activation vector $a^{(0)}$. The result is shown is Table 4.

We observe that treating the image similarity results as

Table 2: Result for graph structured collection, documents not scored

steps	p@10	p@20	r@10	r@20
1	0.232	0.194	0.129	0.195
2	0.233	0.192	0.129	0.195
3	0.22	0.125	0.136	0.204

Table 3: Result for graph structured collection - documents and images not scored

steps	m	p@10	p@20	r@10	r@20
1	10	0.154	0.205	0.046	0.127
2	10	0.154	0.205	0.046	0.127
3	10	0.119	0.165	0.071	0.189

Table 4: Result for graph structured collection, documents scored

steps	p@10	p@20	r@10	r@20
1	0.313	0.229	0.086	0.15
2	0.313	0.229	0.112	0.151
3	0.214	0.196	0.132	0.14

text search results does not culminate in better precision, only 0.011 progress for p@20 and no increase in recall.

4.3.3 Documents scored

We observed worse results in the graph based search in Tables 2 and 3 in comparison with the experiment of static structured data in Table 1. The reason was that we gave the same weight to all neighbours. Though a bias appeared that a (related) document with many images would give less weight to its images (neighbours in the graph) rather than the weight a (non-related) document would give to its few included images.

In this experiment we keep the Lucene score results to propagate in the graph. Therefore, we give the scored value of documents, as activation value in $a^{(0)}$; secondly, in order to remove the bias, we give the document energy completely to its related images by giving weight one to all neighbours. Since these images are part of the documents, this relation is a containment and images can receive the same energy of documents in the first transition step. The result is showed in Table 4. We observe that we obtain better precision of 0.313 for p@10 rather than 0.2 in the previous experiment. Comparing Tables 4 and Table 1, we see that in the 1st and 2nd steps the graph model gives the same precision value as in basic model. The reasons is that in the second step no new images are introduced into the set of results.

Going the first two steps in the collection connected to DBpedia gives the same efficiency since we do not see any different images to affect the precision and recall in these two steps traversal.

5. CONCLUSION AND FUTURE WORK

We described the characteristics of our framework – a generic model for multimodal IR. Astera can model different types of data collections via different link types. Astera is able to enrich the modelled connection by extracting inherent information of data objects as facets, and connecting to semantic network. In this paper we tested this model with ImageCLEF 2011 test collection and showed that Astera obtains similar result in the very first steps of graph traversal. Therefore, there is a lot of potential improvements based on the graph.

As future work, different directions will be followed: 1) Intelligent routing: in each step we filter the results based on meeting a threshold of similarity to the query and continue to traverse the graph based on them, not all the nodes seen in each step. 2) Traversing the graph started from test collection documents through DBpedia till we come back to the collection, considering the effect of semantic links paved. 3) Bringing the image textual information nodes (caption, comment and description) into the game: by computing their similarities to the query topics; further investigating semantic links to their included concepts to the DBpedia nodes. 4) Different weighting to data object facet links based on different modality facets of the query

6. 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.-H. 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] A. Popescu, T. Tsikrika, and J. Kludas. Overview of the wikipedia retrieval task at ImageCLEF 2010. In *CLEF*, 2010.

[13] C. Rocha, D. Schwabe, and M. P. Aragao. A hybrid approach for searching in the semantic web. WWW, 2004.

[14] 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.

[15] S. Srinivasan and M. Slaney. A bipartite graph model for associating images and text. In *Workshop on Multimodal Information Retrieval*, 2007.

[16] A. Tonon, G. Demartini, and P. Cudré-Mauroux. Combining inverted indices and structured search for ad-hoc object retrieval. SIGIR, 2012.