Astera - A Generic Model for Semantic Multimodal Information Retrieval

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

Abstract

Finding useful information from large multimodal document collections such as the WWW is one of the major challenges of Information Retrieval (IR). The many sources of information now available - text, images, audio, video and more - increases the need for multimodal search. Particularly important is also the recognition, that each information item is inherently multimodal (i.e. has aspects in its information character that stem from different modalities) and forms part of a networked set of related information items. In this paper we propose a graph-based model for multimodal information retrieval based on a faceted view of information objects. For retrieval purposes, we consider both relatedness and similarity relations between objects.

1 Introduction

Searching for text, images and audio is common now in Web search and digital libraries. When a user searches for a topic using search engines like Bing and Yahoo, the default category is document. If the user aims to search for other modalities, such as images or videos, she defines it explicitly. However, a user may prefer to see the information in different modalities in the first search and it may happen that she changes the modalities and searches again to find what is more related for her query. Recently we observe a change in this direction at major search engines (i.e. showing a combination of text, image and video in the first page of results, whenever it is considered relevant) further demonstrating the need for a true multimodal system. The limits of current approaches, as observed in these search engines, are the use of essentially one modality to retrieve others (i.e. the use of text features only in retrieving images or videos).

Multimodal IR is generally understood as the combination of text, image, video and sound in information retrieval. In our case, we prefer to generalize this idea, and see Multimodal IR as based on the notion of *facet*. This allows considering a document under several points of view, each one being associated to a possible space For instance, text documents have primarily a textual facet, but also others such as stylistic/layout facets (covered partially by image features), may contain images, or have time/versioning aspects (recency of information). Another example is music files which primarily have audio facets (comprising several actual feature spaces/subfacets such as melodic, rhythmic, chords, voice) but also other facets such as lyrics (as detected from the audio voice), time, genre, etc.

In: M. Lupu, M. Salampasis, N. Fuhr, A. Hanbury, B. Larsen, H. Strindberg (eds.): Proceedings of the Integrating IR technologies for Professional Search Workshop, Moscow, Russia, 24-March-2013, published at http://ceur-ws.org

Furthermore, going beyond the document itself, in modern IR settings, documents are usually not isolated objects: instead, they are frequently connected to other objects, via hyperlinks or meta-data [MCYN06]. Information objects are connected to other information objects, and they provide mutual information on each other, forming a background information model that may be used explicitly. Sometimes this information link is explicit as related information (e.g. a music file and a singer) resulting in a network of related objects; sometimes it is inherent in the information object, e.g. similar pitch histogram of two music files.

There are numerous works in recent years addressing different challenges in multimodal IR. Most of related work try to improve the result relevancy by including different modalities, or focus on ranking issues. Few have worked on addressing different modalities from the very beginning in the search procedure. In this paper, we propose an integrated model for semantic multimodal information retrieval which considers both related and similar objects in the retrieval procedure. Moreover, we employ a faceted view to information objects that enlightens different characteristics of an object, enabling comprehensive and in-depth search.

The rest of the paper is organised as follows: We present the related work in Section 2, followed by the description of our proposed data model in Section 3. We continue with the search procedure in Section 4, and a short summary of the proposed model is provided in Section 5.

2 Related Work

There are many efforts in combining textual and visual modalities. Srinivasan and Slaney [SS07] have improved their performance by adding content based information retrieval, in addition to image characteristics, as visual information. They use a model based on random walks on bipartite graphs of joint modelling of images and textual content.

The combination of both textual and visual features for cross-language image retrieval is addressed by Cheng et al. [CYK+05], who suggest two interactive retrieval procedures. One incorporates a relevance feedback mechanism based on textual information while the second one, combines textual and image information to help users find a target image. Hwang and Grauman have also explored ranking object importance in static images, learning what people mention first from human-annotated tags [HG10].

One of the ideas in the issue of query formulation for multimodal IR is to integrate different modalities to initialize the query. Hubert and Mothe [HM09] suggest a combination of ontology browsing and keyword-based querying. Combining these two modes enables users to complement their queries with keywords for which they do not identify corresponding categories.

Considering the graph-nature of our data model, we look principally at works in the semantic web area. We are taking advantage of the whole semantic web and introduce another feature of similarity checking. Semantic web search is keyword based and there are works on generating adequate interpretations of user queries [SAN+11]. In our model, in addition to including keywords, we consider similarity computation in searching for an object of information. We generalize the query and provide a list of the highly related neighbours to the user, rather than only giving the exact response.

The most related work to our own is the I-Search project, which is a multimodal search engine [LARD12]. They propose a multimodality relation 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. However, they do not consider semantic relation between objects (e.g. a dog and a cat object), nor the importance of these relations in answering the user's query.

3 Graph of Information Objects

We define a model to represent information objects and their relationships, together with a general framework for computing similarity. As shown in Figure 1, we see the information objects as a graph $G = (V, E)$. Each object in this graph has a number of facets. The object modalities could be text, image, audio or video.

For each object, the information can be divided in four categories that applies on different relation types an object holds with neighbors. We formally define relation type $R(e)$ of an edge e, taking one of the four values $R(e) \in \{\alpha, \beta, \gamma, \delta\}$. These types are described as below:

- α: *Related*; this is the *relatedness* relation type and is similar to the relations existing in Semantic Web. For instance a music file object is related to a singer object.
- β: *IsPartof/HasPart*; it is used for showing relation between objects which are part of another object, e.g. an image in a document.

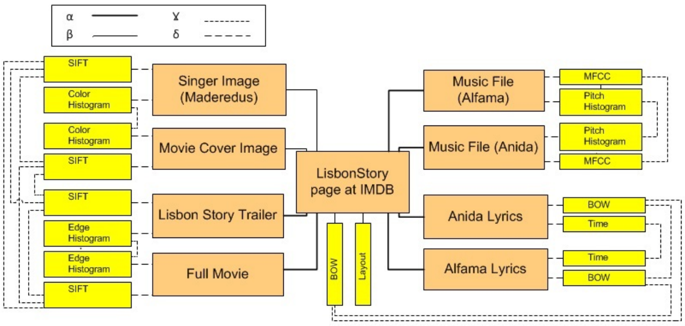

Figure 1: This figure shows a part of related objects of *LisbonStory* movie as a concrete example. Different types of edges ($\alpha, \beta, \gamma, \delta$) are shown between nodes.

- γ: *Similar*; used to show the similarity between objects from the same modality and the same type, e.g. two music files.
- δ: *Inherent/facet relationship*; this type consists of different views to an object, e.g. statistical facet, visual facet, feature facet or genre facet of a piece of music.

An example of mapping this model to a real example is shown in Figure 1. It is about the information related to the movie *Lisbon Story*. As it is shown, the object *LisbonStory page at IMDB* has α relation type with *Music File (Anida)*, *Music File (Alfama)*, *Andia Lyrics*, *Alfama Lyrics*, *Lisbon Story Trailer* and *Full Movie* objects. It has β relation with *Singer Image (Maderedus)* and *Movie cover image* which are the images in the page. Each of these objects have δ relation with their facets, like the relation of *Andia Lyrics* and *BOW* facet. Moreover, we see γ relations between facets of objects. For instance, the *SIFT* feature of the *Full Movie*, *Lisbon Story Trailer* and *Movie cover image* have γ relations to each other.

3.1 Weighting in the Graph

The different types of links described in the previous section may carry with them different weights. We denote the weight of an edge e as $W(e)$. The value of this weight is between 0 and 1, $W(e) = (0, 1]$. For different types of edges, this weight may have different understandings:

- $W(e|R(e) = \alpha) \in (0, 1]$. Since this relation is between two non-homogeneous type objects, we cannot define a weight function. As Crestani [Cre97] mentions, there is no default value for edge weights in spreading activation technique and it is application dependent. Therefore, we assume an initial value of 0.5 for α relations. This may change over time, for instance, based on different relevance feedback techniques.
- $W(e|R(e) = \beta) = 1$. Since this relation is between two objects that are tightly related, one is a part of the other, the value 1 is assigned. The nodes with β relations are extracted from single modal or multi-component objects which are inherently multimodal.
- $W(e|R(e) = \gamma) \in (0, 1]$. This weight is computed by a normalized similarity function between objects of the same type within shared feature spaces. We are aware that normalizing a similarity function is not always obvious, and this is part of the study that this paper starts.
- $W(e|R(e) = \delta) \in (0, 1]$. Similarly to the β edges, the δ relationships have an initial value of 1 because they denote an intrinsic part of the node. The δ edges link an object and its facets in potentially different feature spaces.

3.2 Graph Construction

In this section we explain how we construct the graph with different relation types. The nodes with α relation types are either generated using information extraction techniques from our dataset or extracted from Linked Data [WA11]. The nodes with β relations are created by extracting inherent objects from multimodal objects, e.g. images and text from a PowerPoint presentation. Nodes with γ relationship are generated by computing similarity measures between objects of the same type. Nodes with δ relationships are created in several ways. For instance by feature extraction or by machine learning to learn about, for example, the genre of a music file.

4 Search Procedure

We use spreading activation technique to manage the search procedure, and perform the search on object facets. The weights on edges, which are damping factors, are defined as $df = 1 - w$ in Astera. Therefore higher weighted edges consume less activation energy. After receiving a query, the query facets are extracted. This faceted view of information objects and query enables us to perform search on different characteristics of the objects, resulting in *faceted search.* We hit the graph from N hit points according to query facets and files. In each hit point, parallel multimodal search is conducted basing on spreading activation method. Finally, result collections of different modalities are provided. Our model gives the option of affecting different modalities of the query in search spreading. For instance, if the query consists of both text and music, in searching for each of these modalities, links with neighborhood of the other modality are prioritized.

Astera is capable of representing different retrieval models like vector space, faceted search or multimodal search. Faceted search is directly covered by δ relations, the vector space model can be modelled directly via metrics employed on the facets and gamma relations, with further propagation being set to 0. Multimodal search can be handled both via facets (δ relations) as well as β relations. Semantic search may be modelled by using the α relations. Furthermore, Astera has the potential to answer queries that may not be answerable by VSM or Semantic search individually, but which require a combination of search techniques.

5 Conclusions

In this paper we have introduced a model for multimodal IR with two distinguishing characteristics: one is the idea of faceted view to inherent information encapsulated in objects, which enables us to extract different characteristics of an object to be included in the search procedure. The second is considering both relatedness and similarity relations between objects in the graph model of information objects. The proposed model is domain independent and can be mapped to different domains.

References

[Cre97] F. Crestani. Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, 11(6):453–482, 1997.

[CYK$^+$05] P.C. Cheng, J.Y. Yeh, H.R. Ke, B.C. Chien, and W.P. Yang. Comparison and combination of textual and visual features for interactive cross-language image retrieval. In *Multilingual Information Access for Text, Speech and Images*, pages 919–919. Springer, 2005.

[HG10] S.J. Hwang and K. Grauman. Accounting for the relative importance of objects in image retrieval. In *Proceedings of the British Machine Vision Conference*, pages 1–12, 2010.

[HM09] G. Hubert and J. Mothe. An adaptable search engine for multimodal information retrieval. In *Journal of the American Society for Information Science and Technology*, volume 60, pages 1625–1634. Wiley Online Library, 2009.

[LARD12] M. Lazaridis, A. Axenopoulos, D. Rafailidis, and P. Daras. Multimedia search and retrieval using multimodal annotation propagation and indexing techniques. *Signal Processing: Image Communication*, 2012.

[MCYN06] E. Minkov, W. Cohen, and A. Y. Ng. Contextual search and name disambiguation in email using graphs. In *SIGIR*, pages 27–34, 2006.

[SAN$^+$11] S. Shekarpour, S. Auer, AN 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), 2011 IEEE/WIC/ACM International Conference on Web Intelligence*, volume 1, pages 203–210. IEEE, 2011.

[SS07] S. Srinivasan and M. Slaney. A bipartite graph model for associating images and text. In *IJCAI-2007 Workshop on Multimodal Information Retrieval*, 2007.

[WA11] A. Westerski and Iglesias C. A. Exploiting structured linked data in enterprise knowledge management systems: An idea management case study. In *Enterprise Distributed Object Computing Conference Workshops (EDOCW), 15th IEEE International*, 2011.