Exploiting the Synergy between Different Modalities in Multimodal Information Retrieval

Zhongfei (Mark) Zhang, Ruofei Zhang
Deptartment of Computer Science
State University of New York at Binghamton
USA

Jun Ohya
Global Information and Telecommunication Institute
Waseda University
Japan

Abstract

In this work, we address the issue of exploiting the synergy between different modalities to facilitate Internet information retrieval. It is well-known that information typically exists jointly in different modalities, and this is true for information in the Internet. We intend to demonstrate that exploiting the synergy between the different modalities of the information may enhance the information retrieval efficiency and effectiveness. Specifically, we focus on exploiting the cognitive synergy between text and imagery modalities. Machine learning techniques are used to explicitly exploit this synergy in order to address the effective retrieval of the multimedia information represented in the two modalities. We will demonstrate the proposed methodology through experimental data.

Key words: cognitive synergy, multimodal information retrieval, image retrieval, collateral text, synergistic indexing scheme (SIS), α semantic graph, visible word, abstract word, visual word, inverted file.

1. Introduction

We report an on-going project on multimodal information retrieval through exploiting the cognitive synergy across the different modalities of the information to facilitate an effective Internet information retrieval.

Research on information retrieval in single media types and/or in single modalities has generated extensive literature in well-established communities (e.g., text retrieval [3], image retrieval [9], and video retrieval [2]). However, in many occasions, information does not appear in a single modality, and often information is presented as a combination of different modalities. A typical example is in Internet information retrieval in which a typical Web page often contains textual and imagery as well as other graphic information. On the other hand, although the research on information retrieval in single modalities is well established, it is well known that the retrieval effectiveness is bottlenecked by the notorious semantic gap [9], i.e., almost all the proposed methods in the literature focus on lower-level feature based retrieval, which prevents effective semantic retrieval. Consequently, innovative approaches are necessary to bridge this gap.

Another problem in semantic information retrieval is the *implication problem*. This is related to but not equivalent to the synonymy problem in text retrieval. Taking image retrieval for an example; while there are methods proposed in the literature (e.g., Jing et al [5]) to explicitly "code" the visible "objects" in an image to address the semantic gap problem, these methods still fail to relate an image to an abstract (typically non-visible) concept. For example, in Fig. 1, the image contains visible "objects" such as people, beach, and sky. If a user intends to query any of these concepts, this would become a candidate image. However, if the user intends to query a more abstract concept such as vacation, depending on a specific context, this image might become a related candidate. Yet if the indexing is done only in the image modality or even in a simplistic mapping between image features and words, this image would never be retrieved.

Fig. 1: An image example containing multiple objects.

Based on these considerations, we propose to use multimodal information retrieval to address the semantic gap issue and ultimately to improve the retrieval effectiveness when information retrieval across different modalities is concerned. We believe that there exists synergy between different modalities of the information when the information is presented in many occasions. Taking the advantage of the fact that information often does not appear in single modalities, we can exploit the synergy existing across the different modalities when multimodal information retrieval is concerned. This is also true when only information retrieval in single modalities is required, as in this case we can use the information coexisting in other modalities to exploit the synergy across the modalities

to improve the retrieval effectiveness in the modality in question. An example is Internet image retrieval. Since a typical Web page always contains both text and imagery data, for each image in a Web page, there often exists part of the text in the page explaining the semantic content of the image; even in the case where the Web page is "pure" imagery data, often there is collateral text such as key words to serve as a "caption" to annotate the image. Consequently, exploiting the synergy between the images and their collateral text certainly helps improve the retrieval effectiveness as compared with only focusing on the imagery modality per se in the retrieval. Given an Internet image in Fig. 1, if we only focus on image modality indexing and retrieval, it is very difficult to obtain the semantics of beach and people; the semantics may be obtained by using the collateral text information and exploiting the synergy between the two modalities; even if we had a simplistic mapping between the image features and somewhat "coded" words such as people and beach, this image would still fail to be retrieved if a user queries a more abstract concept such as vacation.

While we believe that exploiting the synergy across different modalities of the information in multimodal information retrieval is new, multimodal information retrieval in general has been investigated in the literature for years. Srihari and Zhang as well as their colleagues investigated image and text retrieval [10]; CMU Informedia project investigated using collateral text from subtitle to retrieve video clips [4]; recently, Li et al investigated using collateral audio to retrieve video [7].

In this project we define the task of *multimodal information retrieval* in a catholic way to incorporate both scenarios where information in different modalities is expected to be retrieved (e.g., retrieve both text and imagery) and where information in only single modalities is expected to be retrieved (e.g., only retrieve imagery). In both scenarios, we use information in different modalities to exploit the synergy across the modalities to address the retrieval problem. Specifically, we focus on the problem of image retrieval when there is collateral text available. This problem appears in many applications such as Internet image retrieval, news image retrieval, and consumer photo retrieval.

2. SIS Based Retrieval

In order to effectively exploit the synergy between imagery and text modalities, we first describe imagery and text indexing, respectively.

For image indexing, we have developed an effective and efficient indexing method in a region based approach resulting in a *Hierarchical Indexing Structure* [11]. In this method, an image is first segmented into a set of regions, and the indexing is then conducted based on the features of the regions. In both the segmentation and the indexing phases, a feature vector consists of color, texture, and shape features. Fuzzy logic is used to accommodate the typical representation impreciseness and segmentation inaccuracy.

In a preliminary research project [11] as an effort to effectively bridge the semantic gap, we have specifically addressed the semantic representation overlap and uncertainty issues. The semantic *overlap* concerns with the representation overlap between two concepts in an image feature space, such as the overlap between "river" and "lake". The semantic *uncertainty* refers to the representation uncertainty w.r.t. a specific concept for a specific image. For example, Fig. 1 can be considered as a "beach" image or as a "people" image, depending on the specific user preference.

To address the two issues, we first use Self Organization Map [6] to generate a visual dictionary in each of the three feature spaces: color, texture, and shape. Each entry of a dictionary is a visual "keywords" quantitatively represented in the imagery feature space using statistical metrics. We use automatic annotation to generate a collection of *visual* words, each of which is grouped by a set of visual "keywords". We manually label all the visual words using textual words based on human cognition.

Then we define a *semantic correlation* for every pair of visual words, based on which, we define the α *semantics graph* as follows.

Definition 2.1: Given the whole set of visual words $D = \{r_1,...,r_n\}$, the semantics correlation c_{ij} defined on the set *D* between any two visual words r_i and r_j, and a real constant $\alpha \in R$, a weighted undirected graph is called α *semantics graph* if it satisfies following constraints: (1) The *node* set of the graph is the symbolic set of all the visual words. (2) There is an *edge* between two nodes iff $c_{ij} \geq \alpha$ and the semantic correlation is the weight of the edge.

The semantic representation overlap and uncertainty issues are resolved through applying fuzzy logic to the α semantics graph. This completes the image modality indexing.

For text indexing, we use the standard *inverted file* approach [8]. The problem, however, is to correctly identify the *relevant* part of the text to a specific image in question as the *collateral* text of the image of *all* the text in a document. This is essentially a layout analysis problem, and it is non-trivial to completely solve for the problem. We use a heuristic approach developed in [10] to identify the collateral text of an image.

Once we have identified the collateral text of an image, we use the standard text processing technique

[8] to extract all the keywords in the collateral text, and then the text indexing is performed based on these keywords using the inverted file approach. This completes the text indexing.

Now we are ready to exploit the synergy between the information of the imagery and text modalities. We first note that there are two types of textual words. A word could be *visible* such that there is a typical visual representation in an image such as people, beach, and building, or could be *abstract* such that the visual representation in an image depends on a specific cognition in a specific context. Examples of the latter case include vacation, work, and appreciation. Clearly, after the manual labeling, all the visual words in the image space are mapped one-to-one to the visible words in the text space.

We propose the *Synergistic Indexing Scheme (SIS)* to explicitly exploit the synergy between the information of the imagery and text modalities. The SIS consists of two spaces: the image space indexed through the hierarchical indexing structure and the α semantics graph, and the text space indexed through the inverted file using a hash table. Fig. 2 illustrates the architecture of the SIS. If a query is given as an image, we can directly use the image indexing based retrieval [11] and there is no point to exploit the synergy. The more interesting problem is to query the database using text and/or imagery.

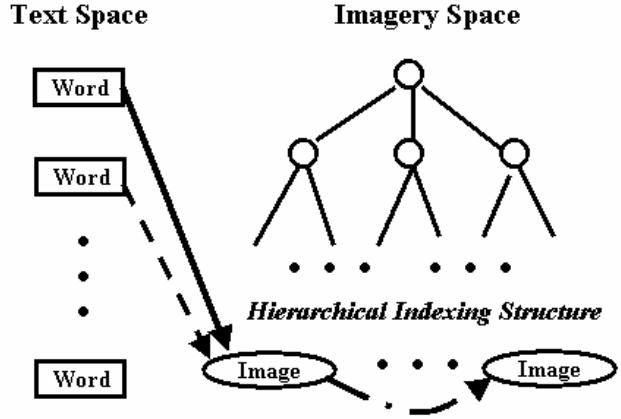

Fig. 2: An illustrative example of the SIS architecture, where the solid arrows are the hard link, the dashed arrows are the soft link, and the line-dot arrows are the relational link in the α semantic graph.

Given a query consisting of text, there are two possibilities. If the query word matches one of the visual words of the image linked through the inverted file in the SIS, the link in the SIS from the word to the image is a *hard link*. This means that this image contains the "information" directly specified by the user, and thus this image definitely needs to be retrieved. Retrieved images falling into this category are called *direct retrieval.* Direct retrieval may be ranked using standard text retrieval methods such as using idf weight [8].

If the query word does not match any of the visual words of the image linked through the inverted file in the SIS, the link in the SIS from the word to the image is a *soft link*. There are two cases.

Case 1: If the word is a visible word, assuming the visual dictionaries are complete covering all the visual words, this query word must be a node in the α semantics graph. We first extend the definition of the α semantics graph to define the following terminologies.

Definition 2.2: Given a specific path $r_1,...,r_k$ between two arbitrary nodes r_1 and r_k in the α semantics graph, the *path correlation* is defined as the sequential product of all the correlations along the path, i.e.,

$$c(r_1,...,r_k) = \prod_{i=1}^{k-1} c_{i,i+1} \qquad (1)$$

Definition 2.3: Given arbitrary two nodes r_1 and r_k in the α semantics graph, the overall correlation between the two nodes is defined as the maximum of all the path correlations between the two nodes.

Let *V* be the set of all the visible words. Based on the above definitions, the extended α semantics graph is essentially a complete graph G(V,E). Given an image database, based on the developed image indexing methods [11] and the above definitions, this extended α semantics graph can be computed offline, and be indexed using standard hashing techniques. Consequently, given a pair of visible words, we can immediately obtain their overall correlation.

Now given the query word *q* and the corresponding linked image in the SIS *I*, let *U* be the set of all the labeled visible words of *I* in the SIS. We define the relational correlation *C(q, I)* as the maximum of all the overall correlations between *q* and every words in U. Thus, given the offline computed extended α semantics graph, *C(q, I)* can be immediately obtained, and the image *I* may be retrieved if *C(q, I)* is above a threshold. Retrieved images falling into this category are called relational retrieval. Relational retrieval may be ranked in terms of *C(q, I).*

Case 2: If the query word is an abstract word, this means that the query word *q* is not in the visible word set *V*, which can be immediately obtained once we have indexed the extended α semantics graph *G(V, E).* This means that we will never be able to describe the concept *q* in the image modality only. Consequently, in this case we must rely on the collateral text to retrieve the image, i.e., we have to trust the collateral text in order to "inference" the semantics of the corresponding linked image, and therefore retrieve the image. Retrieved images falling into this category are called

inferential retrieval. Inferential retrieval may be ranked using standard text retrieval techniques.

At this time the SIS only returns the direct retrieval, relational retrieval, and inferential retrieval separately. How to prioritize the three types of retrieval and how to fuse the three rankings together to generate a single, combined ranking are subject to further investigation. Since the synergy we have exploited between the information of imagery and text modalities is subjective and depends on specific cognitive context, we call this type of synergy as cognitive synergy.

3. Empirical Evaluations

We are now in the process of data collection and implementation of the SIS prototype system. We do not have a full evaluation of the prototype yet. Below we report the evaluation of the α semantics graph based image indexing method.

This part of the evaluation is performed on a general-purpose color image database consisting of 10,000 images from COREL collection. The α semantics graph has 2400 visual "keywords" and has generated 96 visual words. We randomly take 50% of them as the training set to construct the dictionaries and the α semantics graph. To evaluate the image retrieval performance, 1,500 images are randomly selected from the remaining 50% of the collection as the query set. We have invited a group of 5 users to participate the subjective evaluations. The participants consisted of CS graduate students as well as lay-people outside the CS Department. The retrieval relevancy is examined by the users and the retrieval accuracy is the average values across all the query sessions.

In this evaluation we select the α value as the third statistical quartile of all the pair-wise correlations among the visual words, which turns out to be 0.649. To evaluate the contribution of the α semantics graph to the retrieval effectiveness, we have compared the retrieval precision with and without the α semantics graph, and the comparison has shown significant promise of using α semantic graph in delivering semantic retrieval. Considering that it is difficult to design a fair comparison with the existing very few classification-based image retrieval methods, we have compared the average retrieval precision of our method with that of UFM [1], a state-of-the-art CBIR system. The evaluation is shown in Fig. 3. It is clear that both the absolute precision and potential (attenuation trend) of our method are superior to those of UFM.

4. Conclusions

This paper reports an on-going research project on multimodal information retrieval through exploiting the cognitive synergy across the different modalities of the information to facilitate an effective Internet information retrieval. Specifically we focus on Internet image retrieval in the applications where imagery data appear along with collateral text. It is noted that these applications are ubiquitous. We have proposed the Synergistic Indexing Scheme (SIS) to explicitly exploit the synergy between the information of imagery and text modalities. Since the synergy we have exploited between the information of imagery and text modalities is subjective and depends on specific cognitive context, we call this type of synergy as cognitive synergy. We have reported part of the empirical evaluation and are in the process to fully implement the SIS prototype for an extensive evaluation.

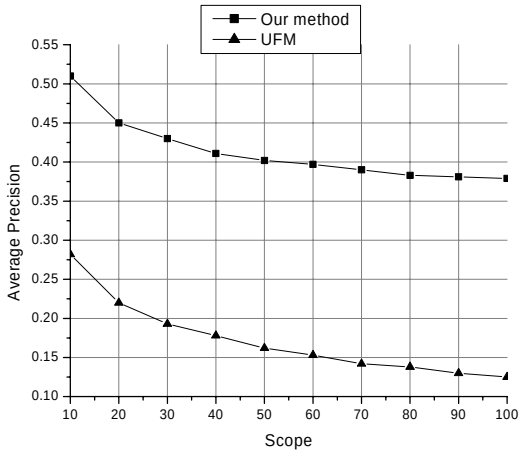

Fig. 3: Comparison between our method and UFM.

References

[1]Y. Chen, J.Z. Wang, A region-based fuzzy feature matching approach to content-based image retrieval, *IEEE T-PAMI*, 24(9), 2002

[2]N. Dimitrova et al, Applications of video-content analysis and retrieval, *IEEE Multimedia*, 9, 2002

[3]C. Faloutsos, D. Oard, A survey of information retrieval and filtering methods, *Tech Report*, Univ. of Maryland, 2002

[4]A. Hauptmann, R. Jin, Visual information retrieval: lessons learned with the Informedia Digital Video Library, *Proc. Int'l Workshop on Digital Communications*, 2002

[5]Feng Jing, et al., An efficient region-based image retrieval framework, *Proc. ACMMM*, 2002

[6]T. Kohonen et al, Self Organization of a massive document collection, *IEEE Trans. on Neural Networks*, 11(3), 2000

[7]D. Li, N. Dimitrova, M. Li, and I.K. Sethi, Multimedia content processing through cross-modal association, ACM MM03, 2003

[8]G. Salton, *Automatic Text Processing*, Addison-Wesley, 1989.

[9]A.W.M. Smeulders et al, Content-based image retrieval at the end of the early years, *IEEE T-PAMI*, 22, 2000

[10]R.K. Srihari, Z. Zhang, A. Rao, Intelligent indexing and semantic retrieval of multimodal documents, 2(2/3), 2000

[11]R. Zhang, Z. Zhang, Addressing CBIR efficiency, effectiveness, and retrieval subjectivity simultaneously, *Proc. ACM Multimedia Indexing and Retrieval Workshop*, 2003