Reachability Analysis of Graph Modelled Collections

Serwah Sabetghadam, Mihai Lupu, Ralf Bierig, and Andreas Rauber

Institute of Software Technology and Interactive Systems
Vienna University of Technology
Vienna, Austria
{sabetghadam,lupu,bierig,rauber}@ifs.tuwien.ac.at

Abstract. This paper is concerned with potential recall in multimodal information retrieval in graph-based models. We provide a framework to leverage individuality and combination of features of different modalities through our formulation of faceted search. We employ a potential recall analysis on a test collection to gain insight on the corpus and further highlight the role of multiple facets, relations between the objects, and semantic links in recall improvement. We conduct the experiments on a multimodal dataset containing approximately 400,000 documents and images. We demonstrate that leveraging multiple facets increases most notably the recall for very hard topics by up to 316%.

1 Introduction

There is rapid growth of online multimodal content as well as personal data generation in our daily life. This trend creates severe challenges in multimodal information retrieval. Multimodal retrieval is defined as searching for the relevant modality with textual queries (keywords, phrases, or sentences) and/or image examples, music files or video clips. Many approaches have been tested in recent years, ranging from associating image with text search scores to sophisticated fusion of multiple modalities [7,5,12].

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). Structured data is naturally represented by a graph, where nodes denote entities and directed/indirected edges represent the relations between them. Such graphs are heterogeneous, describing different types of objects and links. Connected data poses structured IR as an option for retrieving more relevant data objects.

Previous works [9,4,6] introduced models to leverage both structured and unstructured IR. There, a question arises: Is the graph model conducive to retrieval performance? In this work, we propose an analysis on reachability of relevant objects in a graph modelled data. In our previous works [15,16], we introduced a model that enriches the available data by extracting inherent information of

A. Hanbury et al. (Eds.): ECIR 2015, LNCS 9022, pp. 370–381, 2015.

objects in the form of facets. This has support in the principles of Information Retrieval, most notably in the theory of poly-representation [10]. The aim is to leverage cognitive and functional representations of information objects to improve IR results, but there is currently no understanding of how using different representations of the same objects (what we call here facets) affects the reachability of relevant items.

We showed previously that our model matches the efficiency of non-graph based indexes, while having the potential to exploit different facets for better retrieval [16]. In this work, we illustrate the effect of multiple facets on reachability of relevant nodes in a collection. Further, we enrich the relations in the collection by adding corresponding semantic links from DBpedia. We demonstrate how it helps improving recall for hard and very hard topics. We provide extensive experimental evidence for our conclusions, based on the ImageCLEF 2011 Wikipedia dataset [19].

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. Results are discussed in Section 5, and finally, conclusions and future work are presented in Section 6.

2 Related Work

2.1 Content-Based Retrieval

There are many efforts in multimodal retrieval, e.g. by mining the visual information of images to improve text-based search. Martinent et al. [12] propose to generate automatic document annotations from inter-modal analysis. They consider visual feature vectors and annotation keywords as binary random variables. 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].

I-Search, as a multimodal search engine [11], 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. They do not consider explicit links between information objects. We take advantage of the context through links in the context graph whose nodes represent different modalities in the search set.

2.2 Graph-Based Retrieval

Srinivasan and Slaney [18] 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. Jing et al. [8] employ the PageRank to rerank image search. The hyperlinks between images are based on visual similarity of search results. Yao et al. [20] make a similarity graph of images and aim to find authority nodes as result for image queries. Through this model, both visual content and textual information of the images is explored. The structured search engine NAGA [9], provides the results of a structured (not keyword) query by using subgraph pattern on an Entity-Relationship graph. Rocha et al. [13] use spreading activation for relevance propagation applied to a semantic model of a given domain. 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.

In our model, in addition to similarity links between facets of the same type, we have other types of links like semantic or part-of, which enables the framework to model a collection with diverse relation types between information objects. Further, by extracting inherent information of objects in the form of facets, we provide a framework with higher flexibility to prioritize a specific feature. We will show that our model can effectively integrate multiple facets of different modalities to improve performance.

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 and weightings are discussed in detail in [14]. We briefly repeat them here for completeness of the presentation:

- **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 inversely proportional the number of outgoing α links from u.
- **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, and therefore has default weight to 1.
- **Similarity** (γ): relation between objects with the same modality, e.g. between the same facets of two objects. The weight is the similarity value.
- **Facet** (δ): linking an object to its representation(s). It is a directed edge 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.

3.1 Traversal Method - Spreading Activation

There are different methods to traverse a graph of which random walks and spreading activation are two well-known methods. We proved that these two methods are principally the same [17]. However, spreading activation provides more options to customize the graph traversal. 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: $a^{(t)} = a^{(0)} \cdot W^t$ 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.

Memory Spreading Activation algorithm. In this variation of spreading activation, we propose an input function on received energy to manage the amount of energy spreading in the network. The amount of energy a node receives in each step t, is the sum of the energy of its neighbours. Part of this received energy has been sent two steps before from the same node to its neighbours. We subtract this part from the whole received energy to prevent energy bias near activated nodes. We define the *energy capacity* of nodes as vector sm, which contains the sum of the edge weights for each node. We define the energy capacity of node i as $sm_i = \sum_{j=1}^{n} W_{ij}$ where j goes over the columns for row i. This is the energy it is able to carry to its neighbours. It may be less or more than the energy it has at any point in time, as a function of the weights of its outgoing edges. We denote $M = diag(sm)$ which converts vector sm to the diagonal matrix with the vector values on the diagonal. Here, we define the energy received in each step of t as: $a^{(t)} = a^{(0)}.W^t - a^{(t-2)}.M$. In each step, we deduct the self-energy received by subtracting the multiplication of activation value of two steps before to the energy capacity of this node ($a^{(t-2)}.M$). In the expanded form it is:

$$a^{(t)} = a^{(0)} \sum_{k=0}^{t-1} (-1)^k.W^{t-2k}.M^k \tag{1}$$

3.2 Hybrid Search

We proposed to leverage the combination of faceted search with graph search to find relevant objects [15]. 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, which is based on specific facet indexed results. . 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.

The 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 our model, we use the distance constraint to stop the traversal.

4 Experiment Design

4.1 Data Collection

We applied the ImageCLEF 2011 test collection as a benchmark. ImageCLEF 2011 is based on Wikipedia pages and their associated images. It is a multimodal collection (consisting of 125,828 documents and 237,434 images), and an appropriate choice for testing the rich and diverse set of relations in our model.

Each image in this collection has metadata providing 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). 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), δ relation between information objects and their facets.

4.2 Adding Semantic links

We connect the ImageCLEF 2011 Wikipedia collection to DBPedia through the equivalent pages in DBpedia for each wiki page in the collection. The ImageClef2011 Wikipedia collection uses the ImageCLEF 2010 Wikipedia collection, which is based on the September 2009 Wikipedia dumps. Therefore we downloaded DBPedia version 3.4 which is based on Wiki dump September 2009.

Among all DBPedia RDF, we only consider those linking two existing documents in our collection. We add α relations between semantically related documents. The result is a more connected, large scale graph. This way, after visiting a document, we follow its neighbours that may be images or other documents connected through semantic links. For instance, document named *Battle of Leyte Gulf*, contained 6 images as neighbours. After adding semantic links, this document connects to 13 other documents in the collection (e.g. *Pacific War* and *World War II*).

In total 55,544 links are added, which is considerable with respect to the number of documents in the collection (125,828). These links are valuable in the sense that they provide a more connected graph of objects.

4.3 Standard Text and Image Search

In the indexed search approach, as first phase of our hybrid search, we use standard 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:

- **Text tf.idf facet:** We utilize default Lucene indexer, based on tf.idf, as text 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 this facet as **R2**.
- **Image textual annotation tf.idf facet (Tags):** We use metadata information of the images (provided by the collection), as image textual facets (Tags). Meta-data XML files of ImageCLEf 2011, 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.

In the second phase, starting from standard indexed results, we conduct the graph search based on spreading activation to the number of t steps.

4.4 Evaluation Method

The aim of these experiments is to obtain an understanding of the collection of how the relevant images are distributed in the graph. We conduct experiments starting from different indexed facets (Text, CEDD, and Tags).

Through these investigations, we want to see how far and up to how much recall we are able to reach in the graph. There are 50 topics in ImageCLEF 2011 Wikipedia collection. We conduct the traversal up to 50 steps for each of these topics. In each step, we check if we visit new related images for that specific topic. Different topics show different recall behaviour as we go further in the graph. In order to interpret these behaviours, we partitioned the results based on the topic categorization done by Tsikrika et .al [19]. They divide the topics to four categories of easy (17 topics), medium (10 topics), hard (16 topics) and very hard (7 topics). They show 10 topics in easy and hard categories in their report which we use in this work.

5 Results and Discussion

In the first part of the experiments we provide an exploratory data analysis over the collection. In the second part we show the effectiveness of our graph model, leveraging different facets on the collection. In the last part, we perform the same experiments on the semantic enhanced collection.

5.1 Relevant Objects Distribution

Figure 1 shows the distribution of relevant nodes in the collection as we start from all three facets (R1,R2 and R3). The x axis is the number of steps we traverse the graph, and y axis is the Id of the query topics we have. In each step we count the number of new related images we visit. Existence of a shape (circle/square/star/triangle) indicates visiting at least a true positive. The size of a shape is the ratio of number-of-related-seen-nodes-in-this-step/number-of-total-related-ones for the specific query topic Id.

We observe the large number of large shapes, in the first steps. It indicates visiting more related images initiating from different facet results.

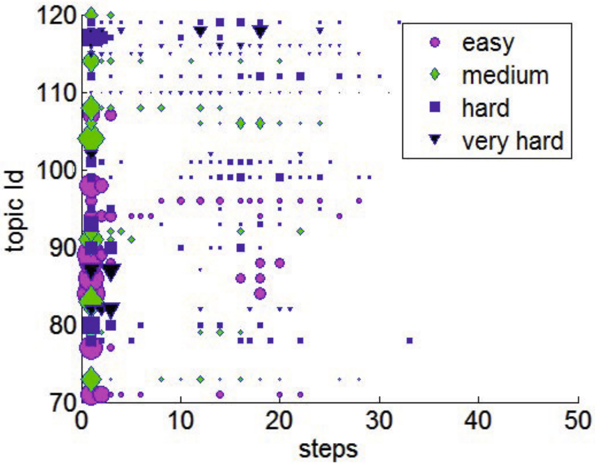


Fig. 1. Relevant node distribution for different categories of topics: easy, medium, hard and very hard

Distribution Per Topic Categories. The distribution of relevant objects for different categories of topics is shown by different shapes in Figure 1. We observe that easy topics points (circles) are mostly at the very beginning steps. For hard and very hard topics (squares and triangles) there are more distributed related nodes as we continue the traversal. They show almost constant increase as we traverse the graph. This observation demonstrates that the distribution of related results for hard and very hard topics is in about 30 steps from the beginning.

5.2 Potential Recall

Here we observe the behaviour of potential recall leveraging different facets.

Different Facet Combinations. We performed the experiment for different combination of facets: R1, R1-R2, R1-R3, and R1-R2-R3 (Figure 2). We observe the changes in the recall values using each combination. The diagram demonstrates that when adding more text (R1-R3) or more image features (R1-R2) we are visiting different objects. In fact R1-R3 results are near to those of R1, while R1-R2 obtains higher recall values, closer to those obtained when using all features (R1-R2-R3). This highlights the importance of different, diverse representations to the data in order to cover all aspects of the relevant objects. The

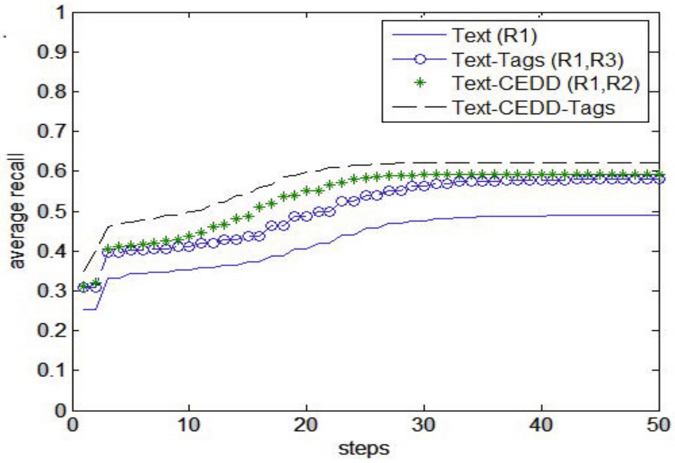


Fig. 2. Average recall for different facets

addition of more textual features, as represented by the meta-data fields (Tags), has produced a lower increase in recall than the addition of CEDD.

We investigate the effect of R1 and R1-R2-R3 facets individually on recall for different categories of topics in next experiments.

Text Facet. In this experiment, we include only R1 results to start search in the graph. Figure 3a shows the average recall for different categories. We observe that easy topics meet 0.66 recall after 27 steps and keeps this value to the 50th step. For medium topics is the same after 25th step with maximum value of 0.51. Hard and very hard topics continue increasing the recall value until 30th step and up to the values of 0.37 and 0.43 respectively. An interesting observation is the behaviour of very hard topics after 3rd step which outpaces hard topics. This demonstrates that as we go farther in the graph we cover higher percentage of recall for very hard topics rather than hard topics. Although we used only Text facet, with the graph modelled collection, we can reach these recall values.

Another observation is the increase rate of average recall in each category. Easy topics show the increase rate of 37.5% (from 0.48 to 0.66), where it is 18.6% for medium topics (from 0.43 to 0.51), 131% for hard topics (from 0.16 to 0.37) and 258% for very hard topics (from 0.12 to 0.43). The values show that hard and very hard topics benefit more than easy topics from the graph structure. While easy and medium topics are apparently answerable by direct query, it is in the hard and very hard topics that the graph model shows most promise.

Further, we observe that recall is increasing up to 30th step and then goes to a plateau for all categories. Two results are obtained from this observation: first is that by conducting the traversal, we can expect increase in recall in the graph to about 30 steps. Because we are still visiting related nodes as we go farther every one or two steps. Second is that after the 30th step we are not visiting relevant images any more, and recall is still less than 0.7 even for easy topics. This shows the disconnectivity of the graph. Our log files show no more node after 40th step for all topics. Therefore, the probability of continuing the traversal and seeing relevant node is zero.

All Facets. We use the R1, R2 and R3 results to start the propagation (Figure 3b). We observe the effect of multiple facets in the beginning steps (1st to 5th) with higher recall values. In addition, the potential recall level can be reached earlier with all facets. We have the same values between 5th and 15th steps here, compared to 15th to 25th step with only Text facet. Further, the average recall has increased to 0.5 for very hard topics (increase rate of 316%).

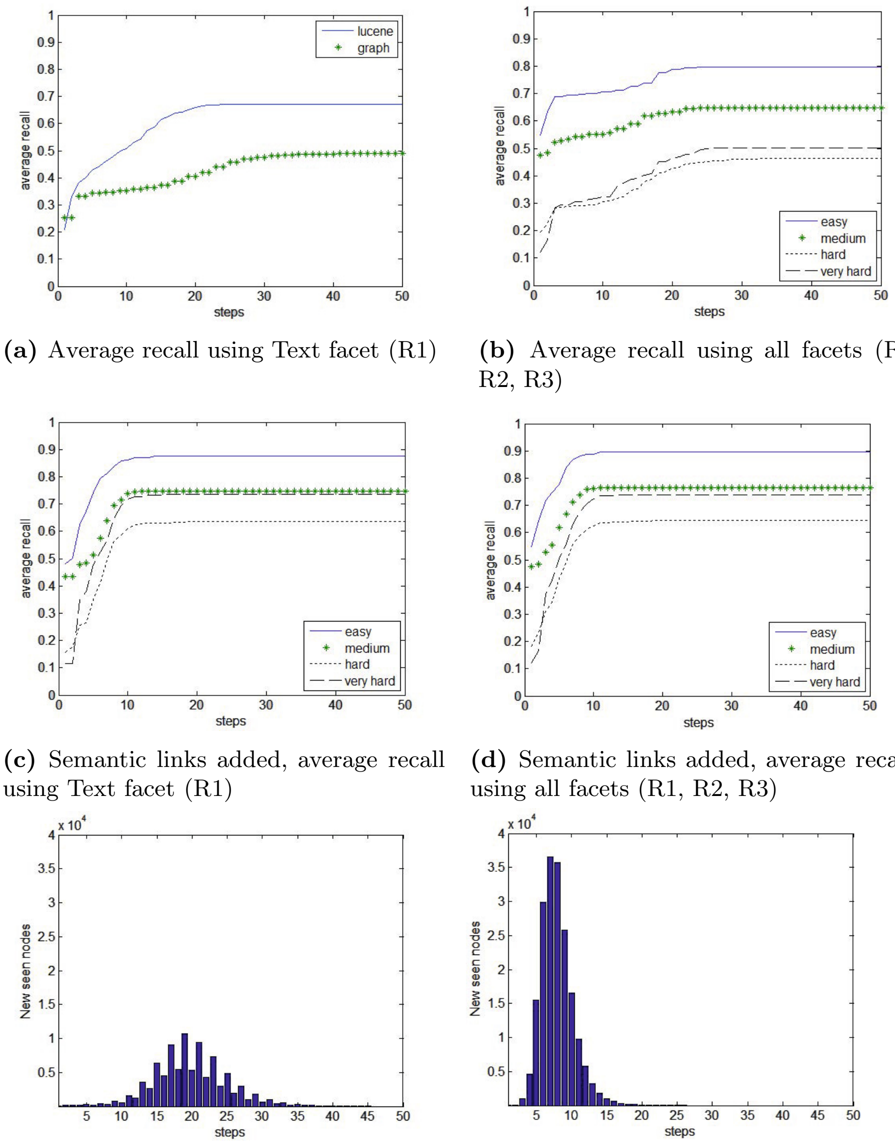
Still Limited View to the Collection. The ImageCLEF 2011 has 363,262 nodes. We counted the number of all seen nodes for different topics. We obtained the average of 93,232 nodes seen starting from all three facets. This illustrates our limited view to this particular collection, by traversing one fourth of the collection size. In addition, the convergence of traversal performance at about 25th-30th step for all topic categories (despite of their different magnitude) is another confirmation to this limited perspective. To tackle this challenge we added semantic links to the collection towards more connectivity.

5.3 Potential Recall - Semantically Enhanced Collection

We perform the same experiments for the collection enhanced with semantic links.

Text Facet. In this experiment, we conduct the test on the enhanced version of the collection including semantic links. It is apparent that we obtain a more connected graph and consequently expect higher recall. We show the reachability result starting from Text facet in Figure 3c. We observe that recall in all categories reaches a plateau in 11th step compared to the graph version without semantic links which was 30 steps. Further, the diagram shows that all categories have a shift in their final value of average recall in comparison to the collection without semantic links: easy topics from %66 to %88, medium from %51 to %75, hard from %37 to %64, and very hard from %43 to %73. In this experiment, hard and very hard topics with 300% and 508%, outpaced other categories.

All Facets. By starting from R1,R2 and R3 results, we reach approximately the same with R1 experiment for different categories (Figure 3d) after 11 steps. The reason is that we have a highly connected graph, of which where to start to search through does not differ after many steps. However, starting from different facets, affects in the initiating steps (1st to 5th step) leading to steeper slope at the beginning. It is considerable since in steps 6, 7 and 8, we are visiting about 30,000 new node in each step. Therefore, for few steps it is worth leveraging different facets, even in highly connected collection.



(a) Average recall using Text facet (R1)

(b) Average recall using all facets (R1, R2, R3)

(c) Semantic links added, average recall using Text facet (R1)

(d) Semantic links added, average recall using all facets (R1, R2, R3)

(e) Number of new seen nodes per step in the collection

(f) Number of new seen nodes per step, semantic links added

Fig. 3. 3a, 3b, 3c, and 3d show recall under different conditions. 3e and 3f show the number of new nodes visited in each step of traversal.

Number of Nodes Seen in Steps. Figure 3e shows the average number of new seen nodes for all topics in each step. We observe that it starts to increase after 11th step up to 30th step to the total size of 93,330 nodes (about 25% of the collection size). The oscillation of the seen nodes in even steps is because of seeing documents in even steps and seeing images in odd steps. The number of images are more than twice of documents in the collection.

The same analysis on the collection containing semantic links demonstrates that the number of nodes are mainly increasing in the first steps up to 11th steps (Figure 3f), to the total size of 188,830 node (about 50% of the collection size). This observation indicates lower number of steps needed to traverse the reachable nodes with semantic links. Further, we touch half of the collection due to more connected collection, leading to visiting more relevant nodes. However, it challenges the precision. Since we visit new nodes in the scale of thousands including few related nodes about 0,001 of the nodes.

6 Conclusion

We presented experiments on the reachability of relevant objects in a graph modelled collection. We compared a graph model where data objects had a set of facets based on their inherent features with a graph model where data objects are additionally connected by semantic links. The results are summarized as below:

- Adding semantic links boosts the potential recall, especially for hard and very hard topic by 300% and 508%.
- Leveraging multiple facets, we saved at least 10 steps to reach the same potential recall compared to using only one facet. Further it increased recall for very hard topics by up to 258%.
- Leveraging semantic links, potential recall reached a plateau in 11 steps. This saved at least 19 steps compared to the traversal without semantic links.
- We demonstrated the effect of different facets leading to visiting different parts of the collection. This reinforces the importance of the poly-representation idea to touch the relevant objects.

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)

Acknowledgments. This research was partly funded by the Austrian Science Fund (FWF) project numbers P25905-N23 (ADmIRE) and I1094-N23 (MUCKE, under the CHIST-ERA Program for Transnational Research Projects).

References

1. Berber, T., Vahid, A.H., Ozturkmenoglu, O., Hamed, R.G., Alpkocak, A.: Demir at imageclefwiki 2011: Evaluating different weighting schemes in information retrieval. In: CLEF (2011)
2. Berthold, M.R., Brandes, U., Kotter, T., Mader, M., Nagel, U., Thiel, K.: Pure spreading activation is pointless. In: CIKM 2009 (2009)
3. Crestani, F.: Application of spreading activation techniques in information retrieval. Artificial Intelligence Review 11 (1997)
4. Delbru, R., Toupikov, N., Catasta, M., Tummarello, G.: A node indexing scheme for web entity retrieval. In: Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) ESWC 2010, Part II. LNCS, vol. 6089, pp. 240–256. Springer, Heidelberg (2010)
5. Duan, L., Li, W., Tsang, I.W., Xu, D.: Improving web image search by bag-based reranking. IEEE Transactions on Image Processing 20(11) (2011)
6. Elbassuoni, S., Blanco, R.: Keyword search over RDF graphs. In: CIKM (2011)
7. Fergus, R., Fei-Fei, L., Perona, P., Zisserman, A.: Learning object categories from google's image search. In: Proc. of Intl. Conf. on Computer Vision (2005)
8. Jing, Y., Baluja, S.: Visualrank: Applying pagerank to large-scale image search. IEEE Trans. Pattern Anal. Mach. Intell. (2008)
9. Kasneci, G., Suchanek, F., Ifrim, G., Ramanath, M., Weikum, G.: Naga: Searching and ranking knowledge. In: ICDE (2008)
10. Larsen, B., Ingwersen, P., Kekäläinen, J.: The polyrepresentation continuum in ir. In: Proc. of IIiX (2006)
11. Lazaridis, M., Axenopoulos, A., Rafailidis, D., Daras, P.: Multimedia search and retrieval using multimodal annotation propagation and indexing techniques. Signal Processing: Image Comm. (2012)
12. Martinet, J., Satoh, S.: An information theoretic approach for automatic document annotation from intermodal analysis. In: Workshop on Multimodal Information Retrieval (2007)
13. Rocha, C., Schwabe, D., Aragao, M.P.: A hybrid approach for searching in the semantic web. In: Proc. of WWW (2004)
14. Sabetghadam, S., Lupu, M., Rauber, A.: Astera - a generic model for multimodal information retrieval. In: Integrating IR Tech. for Prof. Search Workshop (2013)
15. Sabetghadam, S., Lupu, M., Rauber, A.: A combined approach of structured and non-structured IR in multimodal domain. In: ICMR (2014)
16. Sabetghadam, S., Bierig, R., Rauber, A.: A hybrid approach for multi-faceted IR in multimodal domain. In: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., Toms, E. (eds.) CLEF 2014. LNCS, vol. 8685, pp. 86–97. Springer, Heidelberg (2014)
17. Sabetghadam, S., Lupu, M., Rauber, A.: Which one to choose: Random walk or spreading activation. In: Lamas, D., Buitelaar, P. (eds.) IRFC 2014. LNCS, vol. 8849, pp. 112–119. Springer, Heidelberg (2014)
18. Srinivasan, S., Slaney, M.: A bipartite graph model for associating images and text. In: Workshop on Multimodal Information Retrieval (2007)
19. Tsikrika, T., Popescu, A., Kludas, J.: Overview of the wikipedia image retrieval task at imageclef 2011. In: CLEF (2011)
20. Yao, T., Mei, T., Ngo, C.-W.: Co-reranking by mutual reinforcement for image search. In: Proc. of CIVR (2010)

Main Core Retention on Graph-of-Words for Single-Document Keyword Extraction

François Rousseau and Michalis Vazirgiannis

LIX, École Polytechnique, France

Abstract. In this paper, we apply the concept of *k-core* on the *graph-of-words* representation of text for single-document keyword extraction, retaining only the nodes from the main core as representative terms. This approach takes better into account proximity between keywords and variability in the number of extracted keywords through the selection of more *cohesive* subsets of nodes than with existing graph-based approaches solely based on *centrality*. Experiments on two standard datasets show statistically significant improvements in F1-score and AUC of precision/recall curve compared to baseline results, in particular when weighting the edges of the graph with the number of co-occurrences. To the best of our knowledge, this is the first application of graph degeneracy to natural language processing and information retrieval.

Keywords: single-document keyword extraction, graph representation of text, weighted graph-of-words, k-core decomposition, degeneracy.

1 Introduction

Keywords have become ubiquitous in our everyday life, from looking up information on the Web via a search engine bar to online ads matching the content we are currently browsing. Researchers use them when they write a paper for better indexing as well as when they consult or review one to get a gist of its content before reading it. Traditionally, keywords have been manually chosen by the authors but the explosion of the number of available textual contents made the process too time-consuming and costly. Keyword extraction as an automated process then naturally emerged as a research issue to satisfy that need.

A graph-of-words is a syntactic graph that encodes co-occurrences of terms as opposed to the traditional bag-of-words and state-of-the-art approaches in keyword extraction proposed to apply PageRank and HITS on it to extract its most salient nodes. In our work, we capitalize on the k-core concept to propose a novel approach that takes better into account proximity between keywords and variability in the number of extracted keywords through the selection of more cohesive subsets of vertices. The proposed approach presents some significant advantages: (1) it is totally *unsupervised* as it does not need any training corpus; (2) it is *corpus-independent* as it does not rely on any collection-wide statistics such as IDF and thus can be applied on any text out of the box; (3) it *scales* to any document length since the algorithm is linearithmic in the number of unique

A. Hanbury et al. (Eds.): ECIR 2015, LNCS 9022, pp. 382–393, 2015.