

Leveraging Metropolis-Hastings Algorithm on Graph-based Model for Multimodal IR

Serwah Sabetghadam, Mihai Lupu, Andreas Rauber
Institute of Software Technology and Interactive Systems
Vienna University of Technology
Vienna, Austria

ABSTRACT

The velocity of multimodal information shared on web has increased significantly. Many reranking approaches try to improve the performance of multimodal retrieval, however not in the direction of true relevancy of a multimodal object. Metropolis-Hastings (MH) is a method based on Monte Carlo Markov Chain (MCMC) for sampling from a distribution when traditional sampling methods such as transformation or inversion fail. If we assume this probability distribution as true relevancy of documents for an information need, in this paper we explore how leveraging our model with Metropolis-Hastings algorithm may help towards true relevancy in multimodal IR.

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

Theory, Algorithm

Keywords

IR, Multimodal, Graph, Metropolis-Hastings

1. INTRODUCTION

There are many challenges in multimodal information retrieval. Mei et al. [8] have performed a survey on reranking models of multimodal information retrieval. They divide the related work in four categories: 1) *Self-reranking*: includes reranking methods that include data from the original ranking result such as Pseudo-Relevance Feedback or learning a ranking model by giving top ranked documents as positive. 2) *Example-based reranking*: methods to understand the query using accompanying examples. 3) *Crowd reranking*: leverages crowd-sourced knowledge on the web to mine

relevant patterns for a query. 4) *Interactive Reranking*: in this case a user can edit a part of the search results (to delete or to emphasize).

Graph-based methods for reranking are a subset of Self-reranking category, in which a graph oriented search is performed based on relations between objects. Mostly, related work in this area is performed on images/videos with similarity links between them [11, 5]. The use of results from independent modality indexing neglect that data objects are interlinked through different relations. The problem becomes more challenging when the graph is multimodal. During traversal, we may see information objects from different modalities (text, audio, video or image). We propose a model to utilize probabilistic model of IR in multimodal retrieval, with the goal of approaching true relevancy rather than just a reranking. This means that a query may have null result because of lack of any relevant data. According to probability ranking principle in IR, the relevancy of a document to a query is defined as $p(d|q) = \frac{p(q|d)p(d)}{p(q)}$. This requires the probabilities of $p(q)$ and $p(d)$ which are not available. Different ranking models like TF.IDF, BM25 or LM aim to probe the true ranking through different models on $p(q|d)$.

In this paper, we explore the capability of our model to approach probabilistic IR for multimodal retrieval with the help of the MH algorithm. MH is based on MCMC and is used in cases where it is hard to sample from a probability distribution. Assuming the true probability distribution of relevancy of documents to the query as stationary distribution, utilizing MH we make a Markov-chain of documents which results in the same stationary distribution of probabilities. We conduct the experiments on ImageCLEF2011 Wikipedia collection as a multimodal collection.

2. RELATED WORK

There are many efforts in multimodal retrieval in combining textual and visual modalities. Martinent et al. [7] propose to generate automatic document annotations from inter-modal analysis. They consider visual feature vectors and annotation keywords as binary random variables. Jing et al. [6] employ the PageRank to rerank image search. The hyperlinks between images are based on visual similarity of search results. Yao et al. [11] make a similarity graph of images and find authority nodes as result for image queries. Through this model, both visual content and textual information of the images is explored. Hsu et al. [5] leverage context reranking as a random walk over a graph of video stories. The links are based on similarities between different

video stories. The final scoring value is a combination of initial text and stationary distribution scores.

The application of MH method in information retrieval, is limited to search in peer-2-peer networks [3, 1]. Ferreira et al. [3] have designed a protocol for locating a specific object regardless of the topology of the network through uniform sampling from peer-to-peer networks. Zhong et al. [12] use random walks and focus on convergence time for different network sizes. They investigate the probability distribution of visiting nodes. In order to go beyond peer-2-peer networks and apply MH in IR, we need a jumping distribution, i.e. weighted links between nodes. Such links may be similarity/semantic or a mixture of the two. The difficulty, as we will see, is ensuring the stochastic and ergodic nature of the chain.

3. MH ALGORITHM

MH is one of the algorithms based on MCMC to obtain samples from a complex probability distribution $\pi(x)$. The goal is to draw samples from $\pi(x)$ where $\pi(x) = \frac{\tilde{\pi}(x)}{K}$. The normalizing variable K is unknown and hard to compute. Based on the jumping distribution matrix of W , MH algorithm generates a sequence from this distribution as follows:

1. Start with initial value x that $\pi(x) > 0$
2. Using the current x value, sample a candidate point y from $W(x, y)$.
3. The transition probability then is made according to

$$Pr(x, y) = W(x, y)\lambda(x, y) \quad (1)$$

$$\lambda(x, y) = \min\left[\frac{\tilde{\pi}(y).W(y, x)}{\tilde{\pi}(x).W(x, y)}, 1\right] \quad (2)$$

Note that $\lambda(x, y)$ does not require knowledge of the normalizing constant because $\pi(y)/\pi(x)$ drops it out. If it increases the density ($\lambda > 1$), accept y and set the next sample $x_t = y$. Repeat step 3. If it decreases the density, sample u from uniform (0,1). Accept if $\lambda > u$, else reject it.

In order to reach a unique equilibrium state for a Markov-chain, it should be ergodic, satisfying irreducibility (for any state, the probability of getting there given any starting point is more than zero) and aperiodicity (there is no rhythm in which states can be reached given a starting point). There may be different proposal distributions for MH. Two general approaches are [10]: 1) Random walks - the new state y is dependent to the current state x . 2) Independent sample finding - the probability of jumping to point y is chosen from a distribution of interest, independent of the current value. This method is usually used in asymmetric MH. We use the first approach in our work.

4. MODEL REPRESENTATION

We define a graph-based model $G = (V, E)$, in which V is the set of information objects and their facets, and E is the set of edges. By facet we mean inherent feature or representation of an object (e.g., tf.idf facet of a document or edge histogram of an image). Each object may have a number of facets. We define four types of relations. Their characteristics are discussed in detail in [9]. 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_{xy} is made inversely proportional to the α -out-degree of the source node u and $w_{xy} = 1/N_x^{(\alpha)}$.
- **Part-of** (β): a specific type of semantic relation, indicating an object as part of another object, e.g. an image in a document. The weight is 1 because of containment relation as an object part of another one.
- **Similarity** (γ): relation between the facets of two information objects. The weight is the similarity value between the facets.
- **Facet** (δ): linking an object to its representation(s). It is a unidirectional relation from facet to the parent object. Weights are given by perceived information content of features, with respect to the query type.

Our scoring method consists of two steps: 1) In the first step, we perform an initial search with Lucene and/or Lire result based on the facets. This provides us 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.

The model can perform both partial/whole facet retrieval. We may decide to search e.g. only based on query textual or visual facets, or based on all query facets. In practice, we make 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 do not make the result rank list out of top ranked nodes. We initiate their scores in graph nodes and then start propagation. In our model, facet fusion is implicitly calculated by matrix multiplication and final vector computation.

5. MH MAPPED TO IR

We want to achieve a query dependent stationary distribution such that the probability in node x is proportional to the probability that this node is relevant to the query, and at any other node (non-relevant) the probability is zero. This is the $\pi(x)$ distribution from which we cannot directly sample. Instead, we have the $\tilde{\pi}(x)$ which could be a relevance scoring function (e.g. a BM25 score between the data object x_i and the query). MH would formally provide us with a method to sample from the probability distribution, if the approximate probability $\tilde{\pi}$ is properly chosen.

We have the graph of different relations in the adjacency matrix W . Assuming the true relevancy of nodes to the query as $\pi(x)$, we define the $\tilde{\pi}(x)$ as relevance score value function (RSV). A node (M) in the graph may be of any modality: Text (T), Image (I), Audio (A) or Video (V), and the query (Q) may be combination of different modalities. We define the relevance score value function (RSV), as follows:

$$M \in \{T, I, V, A\}$$

$$M = \cup_{i=1}^n M_{f_i}$$

$$Q = \cup_{j=1}^m Q_{f_j}$$

$$l = |\{Q_f | Q_{f_i} = M_{f_i}\}|$$

$$RSV(Q, M) = \sum_{i=1}^l norm(sim(\overline{Q_{f_i}}, \overline{M_{f_i}})) \cdot w_{f_i} \quad (3)$$

where n is the number of facet types of the information object node, m is the number of facet types of the query, sim is the similarity function between two facets, $norm$ is the normalizing function and w_{f_i} is the weight of facet f_i for this query. We compute the similarity (sim) between l number of the same facets of this information object and the query, in which $\overline{Q_{f_i}}$ and $\overline{M_{f_i}}$ are the value of corresponding facets. Usually the value of a facet is in the form of a feature vector. In case of no common facet, the sim function output is zero. Relevancy of an information object to a query should be calculated in accordance to other information objects. For this purpose we compute the similarity of all objects for each query facet and normalize. As we have a multimodal graph and in each step may visit a node with different modality, we require a normalized value to be able to compare the relevancy values.

Different modalities have different facets. Reaching nodes with the same modality of query examples, we have all the facets in common (e.g. an image query and an image node). Visiting nodes with different modality than query examples, we perform similarity for common facets. For instance, if we have an audio object and an image query, we can compare their textual facets (the tf.idf facet of image metadata and tf.idf facet of the audio tags or lyrics).

5.1 MH Constraints in Astera

Irreducibility: To check irreducibility we should prove that our graph is connected. By adding different relations of β , γ and α , we have a connected graph. For this purpose, starting from top ranked results for a sample query we traverse the graph. In each step we visit new neighbours and continue until we see no more new nodes. The number of nodes seen in this traversal was the whole graph size. This observation, even for one query, indicates the connectivity of our graph.

Aperiodicity: Finding nodes from a starting point is not multiple of a number in our graph. We satisfy this constraint by construction.

Stochastic property: According to the weight definition in Astera for β links, the sum of weights on a row may be more than one. However, semantic (α) and/or similarity (γ) links can be used in a normalized form, complying with stochastic property.

Transition Function in Astera According to Metropolis-Hasitngs algorithm, and Eq. 2 we sample from $W(x, y)$ and accept the move with probability $\lambda(x, y)$. This implies on how we define high-order transition probabilities after t steps: $Pr_q^{t+1}(x, y) = \sum_{i=1}^k Pr_q^t(x, z_i)(z_i, y)$ where q is the query, k is the number of common nodes z between x and y , and Pr^t is the transition probability of starting from x and moving t steps further.

Mixing Walsh divides the mixing chains in two categories of **poorly mixing** and **well mixing** chains [10]. To prevent poorly mixing, one usual way is to use Simulated Annealing method with high jumps. Second option is to start with several chains to cover the space to find nodes. Our model follows the second option, as we start from different starting points according to standard search result for each facet.

5.2 Role of MH in Adjusting the Weights

In principle, MH either accepts a jumping density of $W(x, y)$ (when $\lambda > 1$) and keeps the value and moves forward, or modifies the weight with the factor of λ . The new value of this link for next step is $W(x, y) \cdot \lambda$. According to stochastic property, the sum of the weights of links of an edge is 1. In each step, when weights are adjusted by MH, the sum may get lower than 1. In this case the link is accepted with probability of $\lambda < 1$. The decreased value is given as self-transitivity value to the node, indicating staying in this state is preferred than choosing that specific link. Performing this for many steps, loosens the links with less relevant neighbours and keeps the links with increasing relevancy neighbours. This way, MH may modify the weights in the direction of making a Markov chain which reaches to the true probability distribution.

To prevent poorly mixing, we start from different starting points according to standard search result for each facet. These points satisfy the condition of $\tilde{\pi}(x) > 0$ as it is the scored ranked result.

6. EXPERIMENT DESIGN

We applied the ImageCLEF 2011 Wikipedia collection for image retrieval task. Each image has one metadata file that provides information about name, location, one or more associated parent documents in up to three languages (EN, DE and FR), 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.

6.1 Document and Image Facets

In the 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) based on min-max method. Different indexings based on different facets are:

- **Text tf.idf facet:** We utilize default Lucene indexer, based on tf.idf, as text facet.
- **Image textual annotation tf.idf facet (Metadata):** We use metadata information of the images caption, comment and description), as image textual facets.
- **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 [2].

In the second phase, starting from standard indexed results, we conduct the graph search based on MH. In this instantiation of Astera, we use only β links between the documents and images. We investigate adding α and δ link types are in our future works.

6.2 Transition Matrix in Astera

To compute the transition matrix Pr , we need to compute the $\lambda(x, y)$ for each two neighbour nodes to update the weights. In this instantiation of Astera with ImageCLEF

2011 Wikipedia collection, we have images and documents node types. The query topic in this collection is multimodal. It is a combination of keywords and image examples with facet set of $\{tf.idf, CEDD\}$.

Based on any of these facets, we can start traversal in the graph. For example, if we start from similarity with metadata $tf.idf$ results, we will have a set of images as starting points to make the traversal. In this instantiation of Astera, an image object (I) has two facets of $\{tf.idf, CEDD\}$. The common set of facets of l between the query and image is $l = \{tf.idf, CEDD\}$. Each image is connected to at least one parent document (D) through β link. To compute the $Pr(I, D) = W(I, D) \cdot \lambda(I, D)$, we need the λ value, which is:

$$\lambda(I, D) = \left[\frac{RSV(Q, D) \cdot W(D, I)}{RSV(Q, I) \cdot W(I, D)}, 1 \right] \quad (4)$$

where

$$RSV(Q, I) = \frac{norm(sim(\overline{Q_{tf.idf}}, \overline{I_{tf.idf}})) \cdot w_{tf.idf} + norm(sim(\overline{Q_{CEDD}}, \overline{I_{CEDD}})) \cdot w_{CEDD}}{w_{tf.idf} + w_{CEDD}} \quad (5)$$

and

$$RSV(Q, D) = \frac{norm(sim(\overline{Q_{tf.idf}}, \overline{D_{tf.idf}})) \cdot w_{tf.idf}}{w_{tf.idf}} \quad (6)$$

The RSV value is computed based on normalized Lucene and LIRE similarity score for $tf.idf$ and $CEDD$ facet respectively. The w_{CEDD} and $w_{tf.idf}$ are facet weights for this query. For each query, we perform this similarity computation in all three languages, separately for image metadata and documents. We take this value as relevancy value of each image/document for a specific query.

6.3 Experiment Result

We included text $tf.idf$ and metadata $tf.idf$ facets in this experiment. We start with top 20 similar documents and images (as activated nodes) based on these facets for each query, and traverse the graph from these 40 touch points, step by step in parallel. In each step, for node x and its neighbour y , we compute the $\lambda(x, y)$, update the weight and continue to the next neighbour. This is performed in the form of matrix multiplications.

In Markov chain random walks, without MH algorithm, we utilize matrix multiplication to simulate the walk in the graph. The probability distribution after t steps is calculated as $a^t = a^0 \cdot W^t$, where a^0 is the starting scores and a^t is the scores after t steps. However, leveraging MH, the edge weights are affected by λ (Eq. 1). This is a potential problem for computing the updated transition matrix. The reason is that, in each iteration, the matrix W is affected by λ which is a min function - $W \cdot \lambda$ in first iteration and $W \cdot \lambda \cdot \lambda$ in the second iteration. However, Hlynka et al. [4] observed that the transition matrix Pr does not change in further steps. Therefore, we need to compute only once the matrix of $Pr(x, y) = W(x, y) \cdot \lambda(x, y)$ for all nodes, and use this matrix in further multiplications. This makes the MH steps simulation feasible in implementation.

We compute the final score as $a^t = a^0 \cdot Pr^t$ after t steps. This computation is needed for middle steps, since in ideal case the multiplication is performed many times until the matrix converges and in stationary distribution the nodes' probability are independent of starting scores in the graph.

We compare the results with/without using MH algorithm (Tables 1, 2). We did not get better result in our preliminary experiment with MH. The reason is dependency of a jump to the value of $RSV(y)/RSV(x)$. The implemented RSV function for images is based on metadata facet. A large number of images are not retrieved in Lucene result for Metadata facet- we retrieve in the scale of 1000 images for each query, compared to having 274,000 images. We set the minimum value of retrieved scores (0.0001), as RSV value of visited images not in the Lucene results. We have observed that this approach biases a large number of images to very low score, which we assume to be the cause of low precision. Though, further experiments in this direction are needed ¹.

7. CONCLUSION AND DISCUSSION

We presented a graph-based model for multimodal IR leveraging MH algorithm. The graph is enriched by extracted facets of information objects. Different modalities are treated equally thanks to faceted search. We proposed a generic relevancy function based on facet similarity of objects to the query. Leveraging this model, we have a platform, potential to investigate the affect of different facets on performance, and burning in the matrix. We have the opportunity to examine query dependent traversal, as weights in the graph are affected by relevancy of source and target nodes to the query. The preliminary results with MH did not improve the result. Many steps in the graph should be taken until the matrix burns in to the stationary distribution, which is in our future work. However, this experiment brings some issues to discuss: 1) How much the final probability distribution is dependent on the chosen $\tilde{\pi}(x)$? 2) Is MH algorithm on graph-based collections an opportunity to compare the effect of different ranking models? 3) How much expensive is this approach regarding the need of high number of transitions until the matrix burns in? 4) How do we satisfy stochastic property in multimodal graph with heterogeneous relation types? In principle, this property is beyond mathematically summing the weights to 1, but it goes back to the utility of different modalities as neighbours to the user. The difficulty is whether these neighbours are equally useful to the user?

8. REFERENCES

- [1] A. Awan, R. A. Ferreira, S. Jagannathan, and A. Grama. Distributed uniform sampling in unstructured peer-to-peer networks. In *HICSS*, 2006.
- [2] 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.
- [3] R. A. Ferreira, M. Krishna Ramanathan, A. Awan, A. Grama, and S. Jagannathan. Search with probabilistic guarantees in unstructured peer-to-peer networks. In *P2P*, 2005.
- [4] M. Hlynka and M. Cylwa. *Observations on the Metropolis-Hastings Algorithm*. University of Windsor, Department of Mathematics and Statistics, 2009.
- [5] W. H. Hsu, L. S. Kennedy, and S.-F. Chang. Video search reranking through random walk over document-level context graph. *MULTIMEDIA*, 2007.

¹The code of Astera is open-source and available at link: <http://www.ifs.tuwien.ac.at/~sabetghadam/Astera.html>

<i>steps</i>	<i>st</i>	<i>p@10</i>	<i>r@10</i>	<i>p@20</i>	<i>r@20</i>
1	0.297	0.135	0.229	0.158	
2	0.297	0.135	0.229	0.158	
3	0.252	0.123	0.188	0.138	
4	0.224	0.120	0.184	0.134	
5	0.206	0.1148	0.173	0.124	
6	0.182	0.1104	0.156	0.113	
7	0.142	0.106	0.13	0.115	

Table 1: Result for documents without image facets, self-transitivity: 0.9, links: δ, β

<i>steps</i>	<i>st</i>	<i>p@10</i>	<i>r@10</i>	<i>p@20</i>	<i>r@20</i>
1	0.27	0.125	0.151	0.135	
2	0.27	0.125	0.151	0.135	
3	0.23	0.113	0.148	0.1295	
4	0.22	0.1097	0.133	0.1163	
5	0.18	0.1091	0.113	0.1163	
6	0.17	0.107	0.111	0.109	
7	0.14	0.08	0.108	0.087	

Table 2: Result for documents without image facets, self-transitivity: 0.9, links: δ, β

- [6] Y. Jing and S. Baluja. Visualrank: Applying pagerank to large-scale image search. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2008.
- [7] J. Martinet and S. Satoh. An information theoretic approach for automatic document annotation from intermodal analysis. In *Workshop on Multimodal Information Retrieval*, 2007.
- [8] T. Mei, Y. Rui, S. Li, and Q. Tian. Multimedia search reranking: A literature survey. *ACM Computing Surveys (CSUR)*, 2014.
- [9] 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.
- [10] B. Walsh. Markov chain monte carlo and gibbs sampling. 2004.
- [11] T. Yao, T. Mei, and C.-W. Ngo. Co-reranking by mutual reinforcement for image search. CIVR, 2010.
- [12] M. Zhong, K. Shen, and J. Seiferas. The convergence-guaranteed random walk and its applications in peer-to-peer networks. *Computers, IEEE Transactions on*, 2008.