OPTIMIZING STRATEGIES FOR THE EXPLORATION OF SOCIAL NETWORKS AND ASSOCIATED DATA COLLECTIONS

S. Marchand-Maillet, E. Szekely and E. Bruno

Viper group - Department of Computer Science
University of Geneva – 1227 Carouge, Switzerland

ABSTRACT

Multimedia data collections immersed into social networks may be explored from the point of view of varying documents and users characteristics. In this paper, we develop a unified model to embed documents and users into coherent structures from which to extract optimal subsets. The result is the definition of guiding navigation strategies of both the user and document networks, as a complement to classical search operations. An initial interface that may materialize such browsing over documents is demonstrated in the context of Cultural Heritage.

1. INTRODUCTION

Many current information management systems are centered on the notion of a query. This is true over the Web (with all classical Web Search Engines), and for Digital Libraries. In the domain of multimedia, available commercial applications propose rather simple management services whereas research prototypes are also looking at responding to queries. In the most general case, information browsing is designed to supplement search operations. This comes from the fact that the multimedia querying systems largely demonstrate their capabilities using query-based scenario (by Example, by concepts) and these strategies often show limitations, be it in their scalability, their usability or utility or their capabilities or precision. Multimedia search systems are mostly based on content similarity. Hence, to fulfill an information need, the user must express it with respect to relevant (positive) and non-relevant (negative) examples. From there, some form of learning is performed, in order to retrieve the documents that are the most similar to the combination of relevant examples and dissimilar to the combination of non-relevant examples. The question then arises of how to find the initial examples themselves. Researchers have therefore investigated new tools and protocols for the discovery of relevant bootstrapping examples. These tools often take the form of browsing interfaces whose aim is to help the user exploring the information space in order to locate the sought items.

This work is supported by the Swiss NCCR (IM)2 and the EU NoE Petamedia and EU STREP MultiMATCH

The initial query step of most QBE-based systems consists of showing images in random sequential order over a 2D grid. This follows the idea that a random sampling will be representative of the collection content and allow for choosing relevant examples. However, the chance for gathering sufficient relevant examples is low and much effort must be spent in guiding the system towards the relevant region of information space where the sought items may lie. Similarity-based visualization (see *e.g* [1, 2]) organizes images with respect to their perceived similarities. Similarity is mapped onto the notion of distance so that a dimension reduction technique may generate a 2D or 3D space representation where images may be organized. This type of display may be used to capture feedback by letting the user re-organize or validate the displayed images. Specific devices may be used to implement interaction such as interactive tables or multi-touch displays.

A number of similar interfaces have been proposed to apply to the network of users or documents but most browsing operations are based on global hyperlinking (*e.g* Flickr or YouTube pages).

In this paper, we propose a model for the user and document networks (section 2) that unifies navigation as the search of optimal structures within a multigraph (section 3). An example of interface exploiting these structures over documents only is presented in section 4. Based on [3], we aim at adding a user-based browsing functionality to this interface.

2. MULTIDIMENSIONAL DATA MODELLING

We start with a collection $\mathcal{C} = \{d_1, \ldots, d_N\}$ of N multimedia items (text, images, audio, video,...). Traditionally, each document d_i may be represented by a set of features describing the properties of the document for that specific characteristic. In a search and retrieval context, it is expected that mainly discriminant characteristics are considered (*i.e* the characteristics that will make it possible to make each document unique w.r.t a given query). With each of these characteristics is associated a similarity measure computed over the document extracted features. Hence, given $\mathcal{C}$, one may form several similarity matrices $S^{[c]} = \left(s_{ij}^{[c]}\right)$, where the value of $s_{ij}^{[c]}$ indicates the level of similarity between documents d_i and d_j

w.r.t characteristic c.

In our context, we consider each matrix $S^{[c]}$ as a weighted graph connectivity matrix. Since $S^{[c]}$ is symmetric, it represents the connectivity matrix of a complete non-oriented graph where nodes are documents. Collecting all matrices $S^{[c]}\ \forall c$, we may therefore represent our collection as a multigraph acting over the node set $\mathcal{C}$.

This simple similarity-based mapping of the collection provides a useful dimensionless representation over which to act in view over efficient collection exploration. In [4], the High-Dimensional Multimodal Embedding (HDME) is presented as a way to preserve cluster information within multimedia collections. In our context, it forms a useful mapping for projection or dimension selection for visualization. It is also a way of enhancing our Collection Guiding principle proposed in [5].

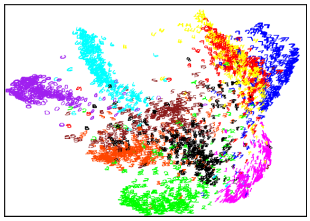

Fig. 1. Global representation of the MNIST handwritten digit image collection after cluster-preserving dimension reduction

Consider now a population $\mathcal{P}$ of K users $\mathcal{P} = \{u_1, \ldots, u_K\}$ interacting with the collection $\mathcal{C}$, thus forming a classical social network over the data. For the sake of generality, we consider that any user u_k may have one or more relationships with a document d_i. For example, user u_k may be the *creator* of document d_i or u_k may have *ranked* the document d_i a certain manner. For each of these possible relationships, we are therefore able to form a matrix $R^{[v]} = \left(r_{ik}^{[v]}\right)$, where the value of $r_{ik}^{[v]}$ indicates the strength (or simply the existence) or relation v between document d_i and user u_k.

Further, users may be associated by inter-relationships (*e.g* the "social graph" , to use the term coined by FaceBook). Classical relationships such as "is a friend of" or "lives nearby" may be quantified for each pair of users. Matrices $P^{[v]} = \left(p_{kl}^{[v]}\right)$ may thus be formed, where the value of $p_{kl}^{[v]}$ indicates the strength of the proximity aspect v between user u_k and user u_l.

In summary, we obtain $(\mathcal{C}, S^{[c]}\ \forall c)$ and $(\mathcal{P}, P^{[v]}\ \forall v)$ as multigraphs acting over document and user node sets and the graph $(\mathcal{C} \cup \mathcal{P}, R^{[v]}\ \forall v)$ as a multi-bipartite graph relating document and user node sets. Figure 2 illustrates this representation.

Now, interestingly, this representation is a base tool for further network analysis and completion.

Graph connectivity analysis of $(\mathcal{P}, P^{[v]}\ \forall v)$ for a given aspect v may tell us about coherence between parts of the

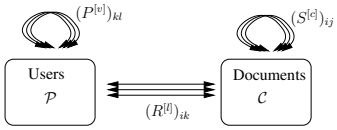

Fig. 2. The proposed graph-based modeling of a social network and associated documents

population. Tools such as minimum vertex- or edge-cuts will indicate particular users or groups that are critical to maintain the connectivity and thus the coherence of the networked information structure.

This structure also allows for its own completion. Similar to what is proposed in [3], user interaction may be captured as one particular bipartite graph $(\mathcal{C} \cup \mathcal{P}, R^{[v]})$ and this information may be mined to enrich either inter-documents similarity $(\mathcal{C}, S^{[c]})$ or inter-user proximity $(\mathcal{P}, P^{[v]})$ to identify a community with specific interests (materialized by the interaction over certain groups of documents). When forming such new relationships, constraints for forming proper distance matrices (1:a) or similarity matrices (1:b) apply:

$$(a)\begin{cases} s_{ij}^{[c]} \geq 0 \\ s_{ij}^{[c]} = s_{ji}^{[c]} \\ s_{ii}^{[c]} = 0 \end{cases} \quad (b)\begin{cases} 0 \leq s_{ij}^{[c]} \leq 1 & \forall i,j \\ s_{ij}^{[c]} = s_{ji}^{[c]} & \forall i,j \\ s_{ii}^{[c]} = 0 & \forall i \end{cases} \qquad (1)$$

Similarly, recommender systems [6] will mine inter-user relationships and inter-document similarity to recommend user-document connections.

3. MULTIDIMENSIONAL DATA EXPLORATION

In this paper, we are interested in defining exploration strategies over social networks (*i.e* joint graphs of documents and users). Based on the above modelling, we map this challenge onto that of defining optimal discrete structures in the multigraphs representing the social network.

3.1. Path-based navigation

The objective is to complement the search paradigm with a navigation facility. We therefore assume that a search tool is used to position a user (called *client* to differentiate from users in the network) at a certain point within our multigraph by selecting a particular user or a particular document. From that point on, the navigation system should enable the client to move within a neighborhood, as defined by the connectivity structure, to explore the vicinity of this position. In other words, we wish to offer the client a view of where to navigate next and this view should be optimized from the information available. Further, this recommendation should be embedded within a global context so as to avoid cycles where the client stays stuck within a loop in the navigation path.

Our graph model is a suitable setup for this optimization. Formally, starting with a matrix $M = (m_{ij})$, where m_{ij} indicated the *cost* of navigating from item x_i to item x_j, we wish to find a column ordering o^* of that matrix that will minimize a certain criteria over the traversal of the items in this order. As a basis, we seek the optimal path that will minimize the global sum of the costs associated to the traversed edges. That is, we seek o^* as the ordering that will minimize the sum of the values above the diagonal so that

$$o^* = \arg\min_{o} \sum_{i \in o} m_{ii+1} \tag{2}$$

The above is equivalent to solving the Symmetric Travelling Salesman Problem (S-TSP) over the complete graph with arc cost m_{ij}. The tour thus forms an optimal discrete structure to explore the complete set of nodes while minimizing the sum of the lengths of the edges traversed during the tour.

In our model, matrices $1 - S^{[c]} = \left(1 - s_{ij}^{[c]}\right)$ and $1 - P^{[v]} = \left(1 - p_{kl}^{[v]}\right)$ follow constraints (1:a) and are therefore suitable inputs for the S-TSP procedure. Figure 3 illustrates the effect of column-reordering on a set distance matrix. The values over the diagonal m_{ii+1} are taken as step costs during the navigation and their overall sum is minimized.

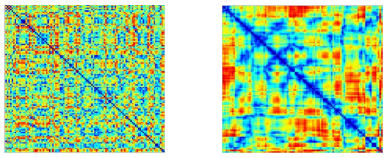

Fig. 3. An optimal reordering applied over a distance matrix (color similarity between 300 images)

Although the S-TSP problem is known to be NP-Complete, effective approximations for solving that problem exist. We use the Lin-Kernighan heuristic [7, 8], known as one of the most efficient approximations. This approximation operates by swapping pairs of sub-tours to make a new shorter tour. At any point of its evolution, this algorithm is therefore able to propose a "best-guess" tour. This enables time-pruning for the computation and it is important to note that, in our context, it is not critical to reach the global optimum. Any deviation from this optimum involving simply swapping items within a close vicinity (which corresponds to the used heuristic) is acceptable.

We therefore construct as many linear paths over the collection of documents as we have characteristics, as S-TSP tours in graphs $(\mathcal{C}, S^{[c]})$. The k-neighborhood of a document d_i is therefore defined as the collection of all k-successors and k-predecessors of d_i over all tours. From there, we can construct an interface that will display this neighborhood and interactively allow proximity moves within the collection.

In parallel, the same strategy may be applied over users and their associated graphs $(\mathcal{P}, P^{[v]})$. The k-neighborhood of a user u_l may be similarly defined over the collection of tours on $\mathcal{P}$.

We therefore obtain an information structure that can be represented as in figure 4.

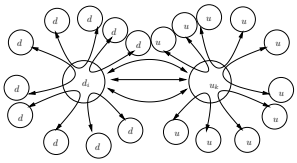

Fig. 4. Defined neighborhood of a particular document or user along defined tours

Both document d_i and user u_k are traversed by their respective paths and are therefore associated with neighbor items (documents and users, respectively). Relationships extracted from graph $(\mathcal{C} \cup \mathcal{P}, R^{[v]})$ enable the transition from d_i to u_k and conversely. Hence, a fully connected navigation flow over the social network of users and documents is created, provided the connectivity in original graphs is suitable.

3.2. Summary-based browsing

The above is based on a proximity-based navigation. That is, from a specific point (user, document, *e.g* provided by a query) the client is able to navigate in the neighborhood of that point along pre-defined paths. From an information space point of view, this means that the client will have difficulty escaping from a given part of the multidimensional presentation space. Only after several navigation step will the client have moved to a part populated with significantly different items.

As a complement to our path-based strategy, we therefore introduce a summary-based browsing that enables the client to "jump" to any place within the navigation space, based on desired criteria. The idea is therefore, based on a given criterion, to create a collection of population summary by sampling the space for diversity and propose to the client a limited but representative subset of the collection or population in question. Here, clustering strategies may be used to obtain a approximation of the space density and then sample accordingly. In order to continue with and take advantage of our graph-based formulation, we use the S-TSP tours to define the collection samples. Given a certain criterion (*i.e* document characteristic c or user inter-relationship v) and a number n of samples to extract, we simply sample the corresponding tour by steps of length

$$L = \frac{\sum m_{ii+1}^*}{n} \tag{3}$$

where m_{ij}^* is the element of the corresponding connectivity matrix $S^{[c]}$ or $P^{[v]}$ reordered according to the presented S-TSP optimization. Equation (3) guarantees that points in the sample set are regularly dispersed within the representation space when considering the geodesic distance (*i.e* sum of edge

lengths instead of Euclidean distance) over the corresponding tour.

4. PROPOSED INTERFACE

In the frame of the MultiMATCH project, we have applied the above principle over a collection of Cultural Heritage digital items composed of images (paintings, historical photographs, pictures,...) annotated with description and metadata. We have defined several browsing dimensions using visual, temporal and textual characteristics. The result is a Web interface presenting a horizontal browsing dimension as its bottom line. From each of these items, a complementary vertical path is displayed. Image size is used to represent the distance from the main focal item displayed at the center of the bottom line. In essence, our browser uses a strategy closes to that implemented in [9].

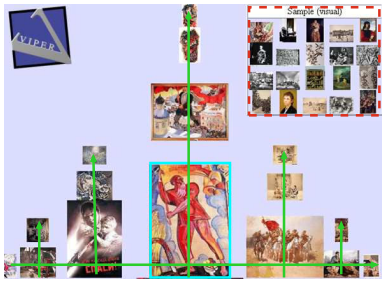

Fig. 5. The proposed browsing interface

Figure 5 shows the interface with the focal point in the bottom blue box. Green vertical and horizontal arrows materialize the paths that may be followed. In the upper right corner (dashed red box), a table displays the summary sample of the collection of items. Clicking on any of these images (but the central one) brings it to the center and updates its context (*i.e* computed neighborhood). Clicking on the central image goes back to the search interface as a complement to the navigation mode and displays the full details (*e.g* metadata) of this particular document. The choice of tours followed along the horizontal and vertical axis is regulated by setting options at each step of the browsing, thus allowing "rotations" around the focal point to display any combination of dimensions.

Concerning the modelling of potential attached social network ($\mathcal{P}$), we are planning to include an interface enabling the browsing of population formed by the creators of the documents, in parallel with this document browsing interface. We have applied the very same principle with different features to browse meeting slide collections with respect to visual slide similarity (to identify reuse of graphical material), textual similarity (to relate presentations by topic and timeline (to simply browse thru the presentation). Again, a social of presentation authors may complement this document browsing tool.

5. CONCLUSION AND DISCUSSION

In this paper, we propose a modelling of data immersed into a social network based on multigraphs. We show how these multigraphs may be the base for defining optimal structures enabling efficient exploitation of the network structure and contained data. In particular, we advocate the use of the S-TSP for organizing a unique navigation path to be followed.

Many graph-based structures are defined by NP-Complete problems. The problem of scalability thus forces proper approximations to be found. We have adapted the classical Concorde implementation of the TSP solver [10] over 70,000 items. Motivated by the fact that in our context, rough approximations are allowed, we are now looking at using a combination of aggregation and TSP solving as a solution to scalability issues.

Our work has been evaluated in all the steps of its engineering (*e.g* similarity computation). However, we still must complete our evaluation by the final usability of the produced interfaces. Initial demonstration sessions with credible tasks show encouraging interests from various classes of users (clients). Only quantitative tests over well-defined tasks and measures will tell us how much these interfaces are actually able to complement classical query-based search operations.

6. REFERENCES

[1] G.P. Nguyen and M. Worring, "Optimization of interactive visual similarity based search," *ACM TOMCCAP*, vol. 4, no. 1, 2008.

[2] Y. Rubner, *Perceptual Metrics for Image Database Navigation*, Ph.D. thesis, Stanford University, 1999.

[3] D. Morrison, E. Bruno, and S. Marchand-Maillet, "Capturing the semantics of user interaction: A review and case study," in *Emergent Web Intelligence*. Springer, 2009, (in press).

[4] E. Szekely, E. Bruno, and S. Marchand-Maillet, "High dimensional multimodal embedding for cluster preservation," Tech. Rep. VGTR:0801, Viper - University of Geneva, 2008.

[5] Stéphane Marchand-Maillet and Éric Bruno, "Collection guiding: A new framework for handling large multimediacollections," in *Audio-visual Content And Information Visualization In Digital Librairies*, Cortona, Italy, 2005.

[6] Thomas Hofmann, "Latent semantic models for collaborative filtering," *ACM Transactions on Information Systems (TOIS)*, vol. 22, no. 1, pp. 89 – 115, 2004.

[7] S. Lin and B. W. Kernighan, "An effective heuristic algorithm for the TSP," *Operations Research*, vol. 21, 1973.

[8] K. Helsgaun, "An effective implementation of the lin-kernighan traveling salesman heuristic," *European Journal of Operational Research*, vol. 126, pp. 106 – 130, 2000.

[9] Scott Craver, Boon-Lock Yeo, and Minerva Yeung, "Multilinearisation data structure for image browsing," in *SPIE Conf. on Storage and Retrieval for Image and Video DBs VII*, 1999.

[10] "The Concorde TSP Solver," `http://www.tsp.gatech.edu/concorde.html`, 2005, (26/01/2009).