Complexity and Algorithms for Composite Retrieval

Sihem Amer-Yahia
CNRS - LIG, France
sihem.amer-yahia@imag.fr

Francesco Bonchi
Yahoo! Research, Spain
bonchi@yahoo-inc.com

Carlos Castillo
Qatar Computing Research Institute - Qatar
chato@acm.org

Esteban Feuerstein
FCEyN, UBA, Argentina
efeuerst@dc.uba.ar

Isabel Méndez-Díaz
FCEyN, UBA, Argentina
imendez@dc.uba.ar

Paula Zabala
FCEyN, UBA, Argentina
pzabala@dc.uba.ar

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Clustering

Keywords

Composite retrieval; Complementarity; Diversity; Maximum Edge Subgraph

1. INTRODUCTION

Online search has become a daily activity and a source of a variety of valuable information, from the finest granularity such as finding the address of a specific restaurant, to more complex tasks like looking for accessories compatible with an iPhone or planning a trip. The latter typically involves running multiple search queries to gather information about different places, reading online reviews to find out about hotels, and checking geographic proximity of places to visit. We refer to this information seeking activity as composite retrieval and propose to organize results into item *bundles* that together constitute an improved exploratory experience over ranked lists.

As a first step towards composite retrieval definition, we need to formalize intuitive desirable properties of item bundles. We distinguish between properties of each bundle in the answer and properties of the answer as a whole.

Consider the case of a user selecting the restaurants to try during a visit to a new city. The user has a limited budget which might be either financial, or simply the number of nights spent in the city. The user prefers suggested restaurants to serve different cuisines. The validity of a bundle of restaurants is given by the *budget constraint* and the *complementarity* of the restaurants in the bundle w.r.t. the cuisine they serve. Other restaurant attributes could be used for defining valid bundles. For example, instead of cuisines, different dress codes could be required to every restaurant in a single bundle. Moreover, in order to provide meaningful bundles, restaurants forming each bundle must be compatible, e.g., close geographically, or liked by similar reviewers. The degree of compatibility of the items forming a bundle defines the quality of the bundle. Intuitively, in the case geographic distance is used, the closer restaurants are from each other, the higher the quality of the bundle they belong to. Similarly, when common reviewers are used as the quality of a bundle, the higher the overlap in similar reviewers between restaurants in the same bundle, the higher the quality of that bundle. Finally, bundles forming an answer set can be generated to cover various geographic areas, or different reviewers segments, thereby producing an answer set of bundles with *diversity.*

Our work is related to result diversification in Web search, database queries, and recommendations, where one aims to achieve a compromise between relevance and result heterogeneity ([3] and references therein). These approaches do not retrieve item bundles. The notion of composite retrieval was proposed with different semantics in recent work ([2] and others). None of these works accounts for diversity.

2. STATEMENT AND COMPLEXITY

Composite retrieval has a wide applicability that goes beyond traditional information retrieval. It is important to note that bundles may be built using the most relevant items to a query thereby making traditional relevance orthogonal to bundle construction. That allows us to define the quality of a bundle, i.e., its score, as a function of pair-wise similarities between its items. As in traditional retrieval, we aim to retrieve highly scoring and also *diverse* bundles. The quality of a collection of k bundles is given by a weighted combination of the quality of each bundle and inter-bundle diversity.

We are given a set of items $\mathcal{I}$. Each item in $\mathcal{I}$ is uniquely identified and has a set of attributes. We assume a *similarity* value $s(u,v)$ in the interval $[0,1]$ for each pair of items $(u,v) \in \mathcal{I} \times \mathcal{I}$. The similarity values $s(u,v)$ may be provided explicitly in the input, or computed implicitly from the representation of the items.

Our goal is to retrieve a set of bundles $\mathcal{S} = \{S_1, \dots, S_k\}$, where a bundle $S_i \in 2^{\mathcal{I}}$ is a set of items that satisfy constraints of *complementarity* and *budget* as expressed in the following definition.

Complementarity: given a property α of the items (e.g., an attribute), no two items in $S_i \in \mathcal{S}$ exhibit the same value for that property: i.e., $\forall u, v \in S_i, u.\alpha \neq v.\alpha$.

Budget: given a set-valued non-negative and monotone function $f : 2^V \rightarrow \mathbb{R}^+$, and given a budget threshold β, we require that $\forall S_i \in \mathcal{S}, f(S_i) \leq \beta$. Typical examples of budget are simply the number of items forming a bundle or an upper-bound on the sum of the costs of items forming the bundle, given a cost attribute.

Composite Retrieval Problem: Given a set of items $\mathcal{I} = \{i_1, \dots, i_n\}$, a pair-wise similarity function $s(u,v)$ for

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

each $(u,v) \in \mathcal{I} \times \mathcal{I}$, a complementarity attribute α, a budget function $f : 2^{\mathcal{I}} \rightarrow \mathbb{R}^+$, a budget threshold β, and an integer k, find a set $\mathcal{S} = \{S_1, \ldots, S_k\}$ of valid bundles that maximizes:

$$\sum_{1\le i\le k} \sum_{u,v\in S_i} \gamma\, s(u,v) + \sum_{1\le i<j\le k} (1-\gamma)(1 - \max_{u\in S_i, v\in S_j} s(u,v))$$

where γ is a user-defined scaling parameter.

The objective function resembles a typical clustering objective, where the total quality of the clustering is expressed as a weighted combination of the quality of single clusters (which in turn is defined as their intra-cluster cohesion) and inter-cluster separation. The latter can be defined as the minimum distance d between an item in one bundle and an item in another one ($d(u,v) = 1 - s(u,v)$). Intra-cluster cohesion reflects cluster quality as a function of the similarity or cohesion between items forming the cluster. Inter-cluster separation reflects answer diversity. Unlike standard clustering, our problem does not seek a total partitioning of items, instead it aims at finding k good groups, that might potentially be small as they are bounded by the budget constraint. Therefore, some items in $\mathcal{I}$ might not belong to any bundle or belong to more than one. Note that our problem definition, by summing over all elements in a bundle, favors larger bundles: as large as possible given the budget constraint.

Complementarity requires no more than one single element of a given kind to belong to a bundle, can be seen as a set of many *cannot-link* constraints typical to constrained clustering. In particular, given the complementarity property α, each item *cannot-link* with all the other items in $\mathcal{I}$ that have the same value for α.

Not surprisingly our problem is hard. We developed two **NP**-hardness proofs, each one highlighting the complexity of one of the arguments of the objective function. Both proofs reduce the well known problem MAXIMUM EDGE SUBGRAPH, which requires to find a set of k nodes, such that the induced subgraph has maximum sum of edge weights.

3. ALGORITHMIC APPROACHES

Given that our problem is **NP**-hard, we turn our attention to approximation algorithms. Going in that direction, we note that MAXIMUM EDGE SUBGRAPH cannot be approximated within constant factors unless **NP** has subexponential time algorithms. We use various heuristics with different approximation ratios proposed in the literature on that problem.

Following the hint of one of our **NP**-hardness proofs, we developed a two-phase approach (*Produce-and-Choose*, or PAC) in which we first produce many valid bundles, and then we choose k among them. For the choosing phase we show an approximation-preserving reduction from MAXIMUM EDGE SUBGRAPH which enables us to adopt heuristics that have been developed in the literature for that problem.

For the task of producing good bundles we observe the similarity between the objective function of our problem, and that of clustering. Following this observation, we devise two ad-hoc clustering algorithms: the first one based on constrained hierarchical clustering, and the second one inspired by *k-nn* clustering. Additonally, we proposed a different method, also suggested by the similarity with a clustering problem. In a first phase items are clustered based on their compatibilities, to form k clusters with good internal cohesion and external separation. This can be done by means of any standard clustering algorithm. Then there is a second phase where we pick a good bundle from each cluster. We refer to this method as *Cluster-and-Pick*, or CAP. Finally, we developed an exact algorithm based on integer linear programming (ILP). We implement a Branch-and-Cut algorithm using CPLEX 12.1, with the addition of a primal heuristic and valid cutting planes specifically derivated for the problem.

4. RESULTS AND FUTURE WORK

We compared experimentally the proposed methods on a large database of user-generated restaurant reviews from Yahoo! Local (38,530 restaurants in 149 US cities), assessing both efficiency and the quality of the results. In terms of efficiency, our heuristics are one order of magnitude faster than the ILP implementation. For instance, CAP has a median running time of 2 seconds, while ILP has a median close to 1 minute (we allow it to run for a maximum of 1 minute, and it often uses that entire amount of time).

In terms of effectiveness, our main finding is that the performance of these methods depends basically on the parameter γ controlling the trade-off between the average score of the bundles and the diversity of the set of bundles. When diversity is highly important (small γ), we obtained the best performance using algorithms of the CAP family. When diversity is less important (large γ), we show that PAC methods that construct good bundles around randomly chosen pivots produce better results. For instance, for a moderate value of budget and a similarity function based on number of reviews in common between two restaurants, the median values of the objective functions are: when $\gamma = 0.1$, CAP$\approx$50 and PAC$\approx$25; but when $\gamma = 0.9$, CAP$\approx$20 and PAC$\approx$50. Extensive experimental results are presented in the poster and **the full version of this work [1]**.

In general, our heuristics produce results comparable to the ILP implementation within a Branch-and-Cut framework. In our experiments, in median the ILP implementation never achieves a value of the objective function of more than twice the value of our best heuristic. In our future work, we plan to explore personalized composite retrieval (e.g., retrieve item bundles that are most compatible with my interests, find the best item bundles including a specific item, find the best bundle compatible with a given item). These new problems bare similarities with item recommendation that account for a user profile, with the added flexibility of querying those recommendations in a stylized fashion. We conjecture that such queries will simplify retrieval complexity while raising an additional challenge of returning results as fast as possible.

5. REFERENCES

[1] S. Amer-Yahia, F. Bonchi, C. Castillo, E. Feuerstein, I. Mendez-Diaz, and P. Zabala. Composite retrieval of diverse and complementary bundles. www.optimization-online.org/DB%5FHTML/2013/02/3785.html. 2013.

[2] M. D. Choudhury, M. Feldman, S. Amer-Yahia, N. Golbandi, R. Lempel, and C. Yu. Automatic construction of travel itineraries using social breadcrumbs. In M. H. Chignell and E. Toms, editors, *HT*, pages 35–44. ACM, 2010.

[3] E. Vee, J. Shanmugasundaram, and S. Amer-Yahia. Efficient computation of diverse query results. *IEEE Data Eng. Bull.*, 32(4):57–64, 2009.