Complexity and Algorithms for Composite Retrieval

Sihem Amer-Yahia
CNRS-LIG
Grenoble, France

Francesco Bonchi
Yahoo! Research
Barcelona, Spain

Carlos Castillo
Qatar Computing
Research Institute.
Doha, Qatar

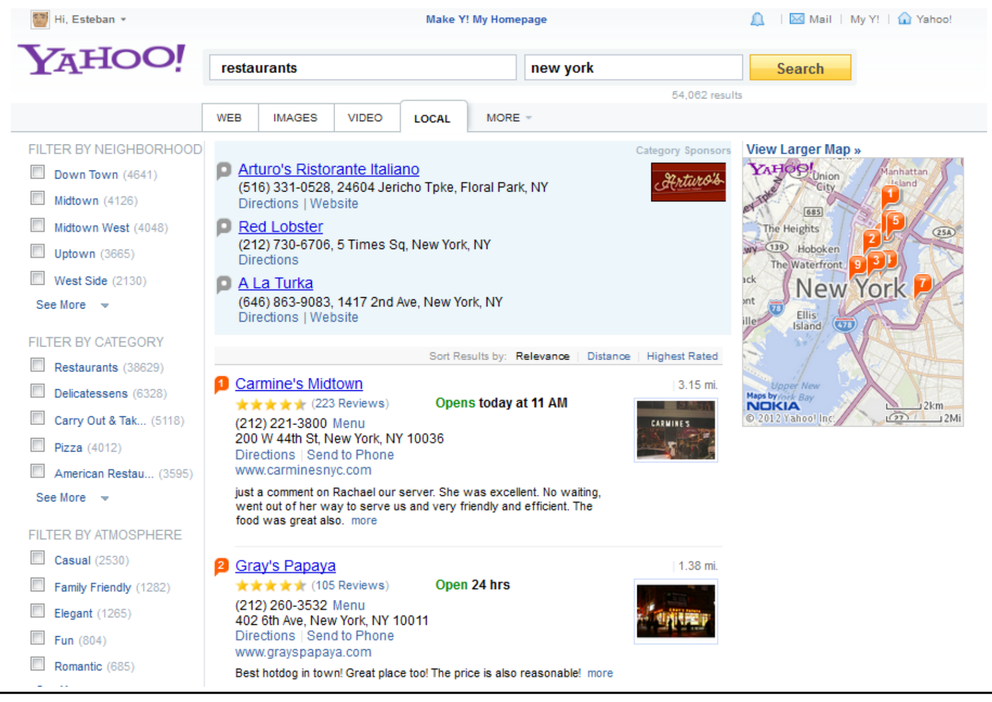
Esteban Feuerstein, Isabel Mendez-Diaz, Paula Zabala
Universidad de Buenos Aires
Argentina

The Problem

Given a set of items, in which each item has a set of attribute values and a cost, and a similarity function for pairs of items, generate a *diverse* set of **composite items** or **bundles**.

Each bundle must be within *budget*, items in each bundle must be *similar*, and two items in the same bundle *cannot have attribute values in common*.

e.g.: Generate itineraries of restaurants/places to visit, instead of a plain ordered list.



Composite retrieval is **NP**-hard, as we prove by reduction from *Maximum Edge Sugraph*. We develop two heuristics:

Produce-and-Choose

Generate candidate bundles, then create a graph of bundles in which nodes are bundles connected by inter-bundle similarities, pick the *k* bundles that minimize inter-bundle similarity using an approximate algorithm for maximum edge subgraph.

Producing bundles #1: C-HAC, constrained hierarchical agglomerative cluster, in which the constraint is that two items with common attributes cannot be together.

Producing bundles #2: BOBO, bundles one-by-one, in which we choose an item as a pivot and greedly build a bundle around that pivot.

Cluster-and-Pick

Also known as CAP. *First phase*: items are clustered by similarity, to form bundles having high intra-similarity.

Second phase: pick a good bundle as a sub-graph of each cluster that respect the complementarity constraint.

Integer Programming

Our *baseline* solution is to state the corresponding integer program, having a number of variables quadratic on the number of items, and solve it using a branch-and-cut method. We let the solver run for at most one minute.

Evaluation

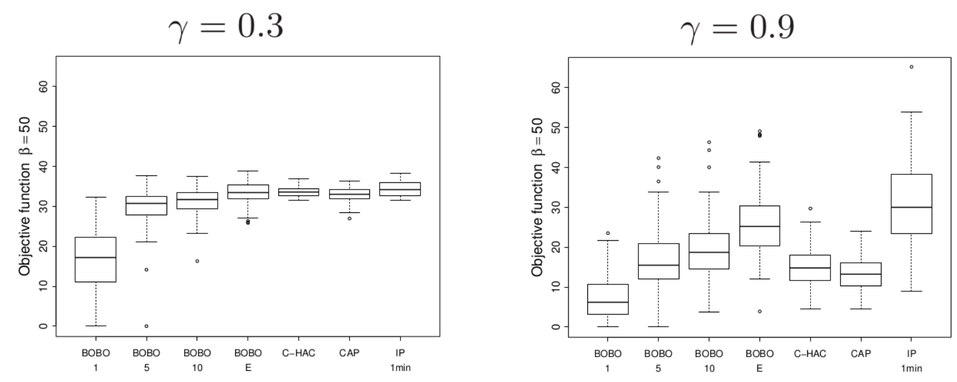
We use a database of 38,530 restaurants in 149 US cities. Each restaurant has average meal prices of $10, $20, or $30. Each restaurant has multiple cuisines (Italian, Chinese, etc.). The similarity between two restaurants is the number of people who have given to both restaurants a positive review.

The goal is to generate 10 bundles for each city, each bundle is a set of recommended restaurants having a sum of cost of $50, $100, or $200 in total.

Experimental results

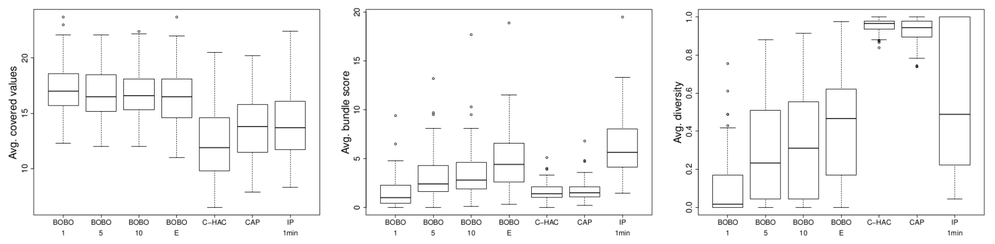
The performance depends on the relative importance of the intra-bundle similarity (γ) vs the inter-bundle similarity (1-γ).

When inter-bundle diversity is very important (small γ), CAP obtains the best performance. Methods such as BOBO that focus on large intra-bundle similarity perform the best when diversity is less important (large γ).



The IP-based method performs the best but it is much slower (avg. 49 seconds vs 2-6 seconds for BOBO and 20 for C-HAC).

The solutions differ in other aspects such as the number of distinct covered attributes or the inter-bundle diversity.



(a) Covered attribute values (b) Bundle score (c) Diversity

For further reading please refer to the poster proceedings or to the the pre-print of the extended version: “Composite Retrieval of Diverse and Complementary Bundles”:

http://optimization-online.org/DB_HTML/2013/02/3785.html

Work done while C. Castillo and S. Amer-Yahia were at Yahoo! Research.