Tutorial

Interactive Exploration of Composite Items

Sihem Amer-Yahia
CNRS, Univ. Grenoble Alpes
Saint Martin D'Hères, France
sihem.amer-yahia@cnrs.fr

Senjuti Basu Roy
New Jersey Institute of Technology
Newark, NJ, USA
senjuti.basuroy@njit.edu

ABSTRACT

Data exploration is seeing a renewed interest in our community. With the rise of big data analytics, this area is growing to encompass not only approaches and algorithms to find the next best data items to explore but also interactivity, i.e. accounting for feedback from the data scientist during the exploration. Interactivity is essential to account for evolving needs during the exploration and also customize the discovery process. In this tutorial, we focus on the interactive exploration of Composite Items (CIs).

CIs address complex information needs and are prevalent in online shopping where products are bundled together to provide discounts, in travel itinerary recommendation where points of interest in a city are combined into a single travel package, and task assignment in crowdsourcing where persoalized micro-tasks are composed and recommended to workers. CI formation is usually expressed as a constrained optimization problem. For instance, in online shopping, package retrieval can retrieve the cheapest smartphones (optimization objective) with compatible accessories (constraints). Similarly, a city tour must be the most popular and conform to a total time and cost budget. A data scientist interested in exploring a variety of CIs has to repeatedly reformulate optimization problems with new constraints and objectives. In this tutorial, we investigate the applicability of interactive data exploration approaches to CI formation.

We will first review CI applications and shapes (15mn). We then discuss three big research questions 60mn): (i) algorithms for CI formation, (ii) modes of exploration for CIs, and (iii) human-in-the-loop CIs. We will conclude with research directions (15mn).

The proposed tutorial is timely. It brings together several related efforts and addresses unsolved questions in the emerging area of human-in-the-loop exploration of complex information needs. The tutorial is relevant to the general area of data science and more specifically to Scalable Analytics, Data Mining, Clustering and Knowledge Discovery, Indexing, Query Processing and Optimization, and Crowdsourcing. The technical topics covered are constrained optimization, ranking semantics, clustering, algorithms, and empirical evaluations.

1 OUTLINE AND SCOPE

1.1 Scope

The tutorial targets theoreticians and practitioners interested in the development of data science applications. It should be of particular interest to an audience who wants to learn about how different domains, such as product recommendation, scientific simulation, or team formation in the social sciences, have been developing their "siloed" definitions of CIs. Tutorial attendees are expected to have basic knowledge in algorithms and data management. Knowledge in constrained optimization is not necessary.

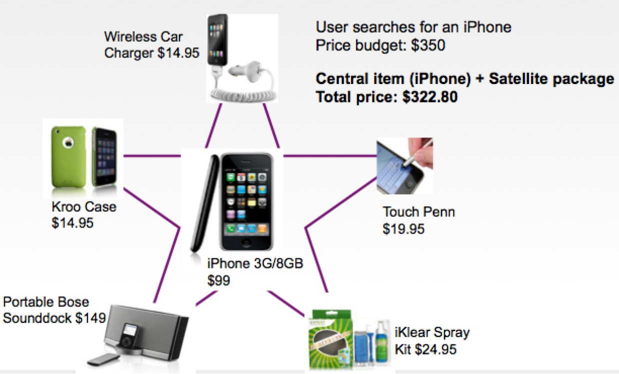

User searches for an iPhone
Price budget: $350
Central item (iPhone) + Satellite package
Total price: $322.80
Wireless Car Charger $14.95
Kroo Case $14.95
Touch Penn $19.95
iPhone 3G/8GB $99
Portable Bose Sounddock $149
iKlear Spray Kit $24.95

Figure 1: A Star CI

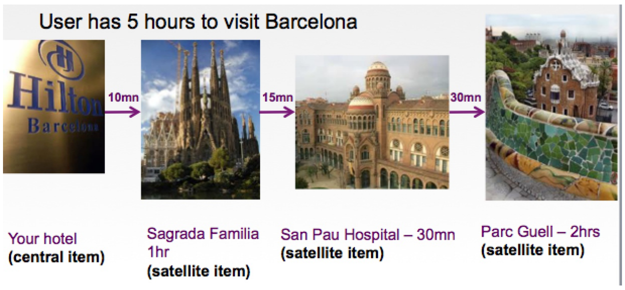


Figure 2: A Chain CI

1.2 Outline

1.2.1 CI Applications and Shapes (15mn). We will begin the tutorial by providing example applications that justify the need for Composite Items (CIs). This part will gather those examples and attempt to unify them. Figures 1, 2, 3 show the many shapes CIs can take in shopping (star shape, where satellite items must be compatible with a central item), traveling (chain shape, where an item must be geographically close to its preceding item), and dining (where items are jointly co-reviewed by the same people), respectively.

1.2.2 Research Questions: (60mn). As our tutorial topic is *building* and *exploring* CIs *interactively*, the second part of the tutorial is organized into three big research questions: (i) algorithms for CI retrieval, (ii) how data exploration is typically done for large-scale datasets and its applicability to exploring CIs, and (iii) what types of user interactivity are common and their applicability to building and exploring CIs interactively. For each question, we will review the state of the art and discuss new research challenges.

Research Question 1: CI Retrieval (20mn). Different CI shapes require the specification of different constraints and optimizations, thereby leading to a no "one-size-fits-all" CI definition. We will discuss why and how the nuances of data, such as type heterogeneity, dimensionality, distribution, or even storage, impact

Published in Proceedings of the 21st International Conference on Extending Database Technology (EDBT), March 26-29, 2018, ISBN 978-3-89318-078-3 on OpenProceedings.org.

Series ISSN: 2367-2005

10.5441/002/edbt.2018.60

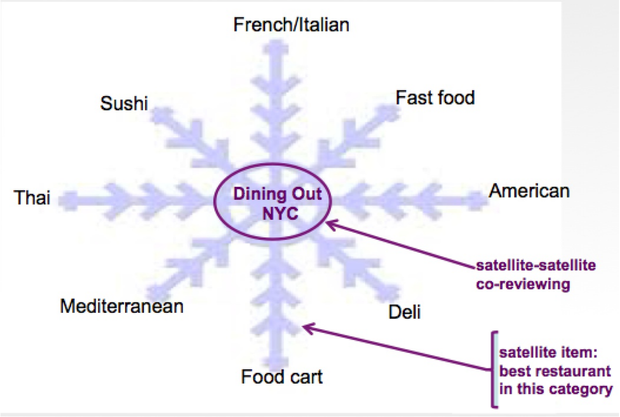
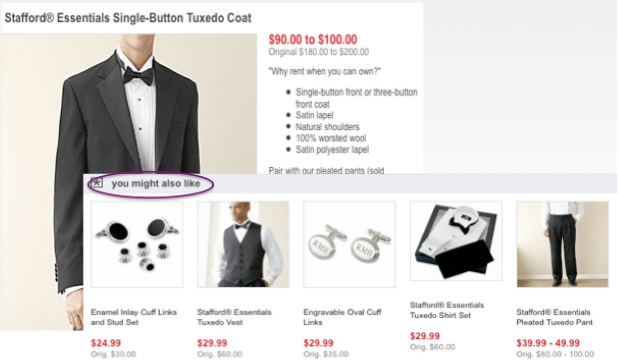

French/Italian
Sushi
Fast food
Thai
Dining Out NYC
American
satellite-satellite co-reviewing
Mediterranean
Deli
Food cart
satellite item: best restaurant in this category

Figure 3: A Snowflake CI



Stafford® Essentials Single-Button Tuxedo Coat

$90.00 to $100.00
Original $180.00 to $200.00

"Why rent when you can own?"

- Single-button front or three-button front coat
- Satin lapel
- Natural shoulders
- 100% worsted wool
- Satin polyester lapel

Pair with our pleated pants (sold

you might also like

Enamel Inlay Cuff Links and Stud Set
$24.99
Orig. $30.00

Stafford® Essentials Tuxedo Vest
$29.99
Orig. $60.00

Engravable Oval Cuff Links
$29.99
Orig. $35.00

Stafford® Essentials Tuxedo Shirt Set
$29.99
Orig. $60.00

Stafford® Essentials Pleated Tuxedo Pant
$39.99 - 49.99
Orig. $80.00 - 100.00

Figure 4: Recommending Satellite Items

the semantics and performance of building different shapes of CIs. For example, the type of aggregation that a climatologist seeks to build a CI that reflects a meaningful "climate model" is significantly different from that of a retail business manager who packages compatible items for product recommendation, or from a crowdsourcing platform that bundles diverse micro-tasks to address workers' bordedom. Using these applications, we will present scenarios where exploration of CIs and interactivity are essential. For instance, in the domain of experiment design, we will show the necessity of leveraging feedback from domain scientists to select a different set of parameters that appropriately capture a scientific simulation process, and represent it as a CI. In the travel world, we will argue that interactivity helps refine one's partial needs and build personalized packages [20, 24]. Similarly, in crowdsourcing, a CI may represent a bundle of diverse tasks that are more exciting to workers than a ranked list of tasks [4].

Research Question 2: CI Exploration (20mn). The aim of this part is to bridge the gap between exploration in emerging data science applications and exploring structured data, and discuss how that can serve CI exploration.

We will first describe approaches that are popularly used among data scientists to explore large-scale datasets. Such data exploration techniques lack well-defined objectives and are mostly done following a trial-and-error approach. Consequently, most of the visualization-based data exploration techniques that data scientists popularly use are ad-hoc and unprincipled.

After that we will discuss data exploration techniques that are more principled and investigated in conjunction with structured databases, especially considering a user input, in the form of queries, example data, or data distributions. Some notable examples in that space relate to faceted search and query expansion techniques [8, 9, 11, 12, 14, 16, 21, 29, 30]. We will also discuss some recent work that investigates exploration and visualization techniques intended to assist users by looking for similar data distributions [5, 22, 23, 26] or in an example-driven exploration [15, 19]. We will argue why these techniques are not directly applicable to CI exploration.

Finally, we will discuss why CI exploration needs to go beyond existing techniques and rely on optimization-guided data exploration as in [18]. We will pose open problems and computational challenges in designing appropriate solutions for exploring CIs.

Research Question 3: Human-in-the-Loop CIs (20mn). The third challenge is how to incorporate human-in-the-loop effectively to enable interactive data exploration.

We will first discuss different types of interactivity that a user is allowed to provide, ranging from binary responses to capturing implicit actions. We will do that in the context of different contexts such as crowdsourcing, recommender systems, experiment design, or machine learning applications that require supervised samples for training models. In conjunction, we will discuss different types of users, such as naive users or domain experts, and investigate what types of challenges, such as expressivity of interaction, bias, and real-time interactions, they incur. We will examine how these questions can be revisited in the context of building CIs and review preliminary work on satellite item recommendation [7, 20] (also see Figure 4), and on adding or deleting specific items in CIs [24] (also see Figure 5). An additional challenge when interacting with CIs is the visual layout. Unlike "flat" items, aiding users in their interactions with CIs, via recommendations or maps for instance, is necessary for a full-fledged exploration.

In a second part, we will describe other types of feedback that are rather non-traditional and only discussed in recent work, such as feedback on metadata rather than on the entire object. We will discuss how such feedback is processed for answering queries, feature selection, or feature engineering in the data science pipeline [10, 17, 32]. We will describe our vision on hybrid approaches that discuss how to leverage sophisticated yet limited human feedback in the computational loop and show their utility for CIs.

1.3 Research Challenges: (15mn)

In this last part of the tutorial, we will summarize and brainstorm our overarching framework to enable optimization-guided data exploration techniques that enable a human-in-the-loop approach. We will discuss what types of applications it will support and conclude by outlining some major challenges in combining CI exploration and interactive CIs.

1.4 Overlap with Previously Presented Tutorials

In WWW 2015, the authors presented a tutorial on composite items in the context of complex crowdsourcing [1]. This tutorial will have a very small overlap with that tutorial in the part that reviews various CI definitions. Other than that, the content of this tutorial was not presented before.

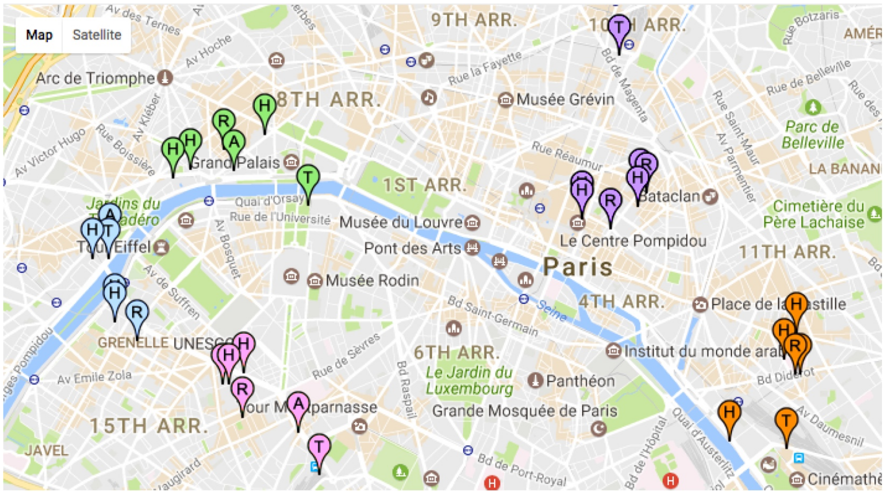


Figure 5: CIs on a Map

2 BIOGRAPHY

The authors published seminal papers on composite item retrieval and data exploration.

Sihem Amer-Yahia, *sihem.amer-yahia@cnrs.fr,* is a Research Director in Grenoble. Her interests are at the intersection of data management and social data exploration. She is the Editor-in-Chief of the VLDB Journal for Europe and Africa and PC co-chair of PVLDB 2018 and WWW Tutorials 2018.

Senjuti Basu Roy, *senjuti.basuroy@njit.edu,* is an Assistant Professor at NJIT. Her research interests lie in the area of data and content management of web and structured data with a focus on exploration, analytics, and algorithms. She is the PC Co-chair of SIGMOD 2018 mentorship track, PC co-chair of VLDB 2018 PhD workshop track.

REFERENCES

[1] Sihem Amer-Yahia and Senjuti Basu Roy. 2015. From Complex Object Exploration to Complex Crowdsourcing.. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 1531–1532.

[2] Sihem Amer-Yahia, Francesco Bonchi, Carlos Castillo, Esteban Feuerstein, Isabel Méndez-Díaz, and Paula Zabala. 2013. Complexity and algorithms for composite retrieval. In *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 79–80.

[3] Sihem Amer-Yahia, Francesco Bonchi, Carlos Castillo, Esteban Feuerstein, Isabel Mendez-Diaz, and Paula Zabala. 2014. Composite retrieval of diverse and complementary bundles. *IEEE Transactions on Knowledge and Data Engineering* 26, 11 (2014), 2662–2675.

[4] Sihem Amer-Yahia, Éric Gaussier, Vincent Leroy, Julien Pilourdault, Ria Mae Borromeo, and Motomichi Toyama. 2016. Task Composition in Crowdsourcing. In *IEEE International Conference on Data Science and Advanced Analytics, DSAA, Montreal, QC, Canada, October 17-19, 2016*. 194–203.

[5] Sihem Amer-Yahia, Sofia Kleisarchaki, Naresh Kumar Kolloju, Laks V. S. Lakshmanan, and Ruben H. Zamar. 2017. Exploring Rated Datasets with Rating Maps. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*. 1411–1419.

[6] Sihem Amer-Yahia, Laks Lakshmanan, and Cong Yu. 2009. Socialscope: Enabling information discovery on social content sites. *arXiv preprint arXiv:0909.2058* (2009).

[7] Senjuti Basu Roy, Sihem Amer-Yahia, Ashish Chawla, Gautam Das, and Cong Yu. 2010. Constructing and exploring composite items. In *ACM SIGMOD*. ACM, 843–854.

[8] Senjuti Basu Roy, Haidong Wang, Gautam Das, Ullas Nambiar, and Mukesh Mohania. 2008. Minimum-effort driven dynamic faceted search in structured databases. In *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 13–22.

[9] Adriane Chapman and HV Jagadish. 2009. Why not?. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*. ACM, 523–534.

[10] Justin Cheng and Michael S Bernstein. 2015. Flock: Hybrid crowd-machine learning classifiers. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 600–611.

[11] Kyriaki Dimitriadou, Olga Papaemmanouil, and Yanlei Diao. 2014. Explore-by-example: An automatic query steering framework for interactive data exploration. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*. ACM, 517–528.

[12] Zhian He and Eric Lo. 2014. Answering why-not questions on top-k queries. *IEEE Transactions on Knowledge and Data Engineering* 26, 6 (2014), 1300–1315.

[13] Arlind Kopliku, Karen Pinel-Sauvagnat, and Mohand Boughanem. 2014. Aggregated search: A new information retrieval paradigm. *ACM Computing Surveys (CSUR)* 46, 3 (2014), 41.

[14] Chengkai Li, Ning Yan, Senjuti B Roy, Lekhendro Lisham, and Gautam Das. 2010. Facetedpedia: dynamic generation of query-dependent faceted interfaces for wikipedia. In *Proceedings of the 19th international conference on World wide web*. ACM, 651–660.

[15] Davide Mottin, Matteo Lissandrini, Yannis Velegrakis, and Themis Palpanas. 2017. New Trends on Exploratory Methods for Data Analytics. *PVLDB* 10, 12 (2017), 1977–1980.

[16] Davide Mottin, Alice Marascu, Senjuti Basu Roy, Gautam Das, Themis Palpanas, and Yannis Velegrakis. 2013. A probabilistic optimization framework for the empty-answer problem. *Proceedings of the VLDB Endowment* 6, 14 (2013), 1762–1773.

[17] Besmira Nushi, Adish Singla, Andreas Krause, and Donald Kossmann. 2016. Learning and Feature Selection under Budget Constraints in Crowdsourcing. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

[18] Behrooz Omidvar-Tehrani, Sihem Amer-Yahia, and Alexandre Termier. 2015. Interactive User Group Analysis. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*. 403–412.

[19] Olga Papaemmanouil, Yanlei Diao, Kyriaki Dimitriadou, and Liping Peng. 2016. Interactive Data Exploration via Machine Learning Models. *IEEE Data Eng. Bull.* 39, 4 (2016), 38–49. http://sites.computer.org/debull/A16dec/p38.pdf

[20] Senjuti Basu Roy, Gautam Das, Sihem Amer-Yahia, and Cong Yu. 2011. Interactive itinerary planning. In *Data Engineering (ICDE), 2011 IEEE 27th International Conference on*. IEEE, 15–26.

[21] Senjuti Basu Roy, Haidong Wang, Ullas Nambiar, Gautam Das, and Mukesh Mohania. 2009. Dynacet: Building dynamic faceted search systems over databases. In *Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on*. IEEE, 1463–1466.

[22] Tarique Siddiqui, John Lee, Albert Kim, Edward Xue, Xiaofo Yu, Sean Zou, Lijin Guo, Changfeng Liu, Chaoran Wang, Karrie Karahalios, et al. 2017. Fast-Forwarding to Desired Visualizations with Zenvisage.. In *CIDR*.

[23] Tarique Siddiqui, John Lee, Albert Kim, Edward Xue, Xiaofo Yu, Sean Zou, Lijin Guo, Changfeng Liu, Chaoran Wang, Karrie Karahalios, and Aditya G. Parameswaran. 2017. Fast-Forwarding to Desired Visualizations with Zenvisage. In *CIDR 2017, 8th Biennial Conference on Innovative Data Systems Research, Chaminade, CA, USA, January 8-11, 2017, Online Proceedings*.

[24] Manish Singh, Ria Mae Borromeo, Anas Hosami, Sihem Amer-Yahia, and Shady Elbassuoni. 2017. Customizing Travel Packages with Interactive Composite Items. In *2017 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2017, Tokyo, Japan, October 16-18, 2017*.

[25] Quoc Trung Tran and Chee-Yong Chan. 2010. How to conquer why-not questions. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. ACM, 15–26.

[26] Manasi Vartak, Samuel Madden, Aditya Parameswaran, and Neoklis Polyzotis. 2014. SEEDB: automatically generating query visualizations. *Proceedings of the VLDB Endowment* 7, 13 (2014), 1581–1584.

[27] Min Xie, Laks VS Lakshmanan, and Peter T Wood. 2010. Breaking out of the box of recommendations: from items to packages. In *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 151–158.

[28] Min Xie, Laks VS Lakshmanan, and Peter T Wood. 2011. Comprec-trip: A composite recommendation system for travel planning. In *IEEE International Conference on Data Engineering (ICDE)*. 1352–1355.

[29] Ning Yan, Chengkai Li, Senjuti B Roy, Rakesh Ramegowda, and Gautam Das. 2010. Facetedpedia: enabling query-dependent faceted search for wikipedia. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 1927–1928.

[30] Sivan Yogev, Haggai Roitman, David Carmel, and Naama Zwerdling. 2012. Towards expressive exploratory search over entity-relationship data. In *Proceedings of the 21st International Conference on World Wide Web*. ACM, 83–92.

[31] Nan Zhang, Chengkai Li, Naeemul Hassan, Sundaresan Rajasekaran, and Gautam Das. 2014. On skyline groups. *IEEE Transactions on Knowledge and Data Engineering* 26, 4 (2014), 942–956.

[32] James Y Zou, Kamalika Chaudhuri, and Adam Tauman Kalai. 2015. Crowdsourcing feature discovery via adaptively chosen comparisons. *arXiv preprint arXiv:1504.00064* (2015).