

# Report on RecSys 2015 Workshop on New Trends in Content-Based Recommender Systems

Toine Bogers<sup>1</sup>      Marijn Koolen<sup>2</sup>  
toine@hum.aau.dk    marijn.koolen@uva.nl  
<sup>1</sup>Aalborg University Copenhagen, Denmark  
<sup>2</sup>University of Amsterdam, Netherlands

## Abstract

This article reports on the *CBRecSys 2015* workshop, the second edition of the workshop on new trends in content-based recommender systems, co-located with RecSys 2015 in Vienna, Austria. Content-based recommendation has been applied successfully in many different domains, but it has not seen the same level of attention as collaborative filtering techniques have. Nevertheless, there are many recommendation domains and applications where content and metadata play a key role, either in addition to or instead of ratings and implicit usage data. The CBRecSys workshop series provides a dedicated venue for work dedicated to all aspects of content-based recommender systems.

## 1 Motivation and Goals

While content-based recommendation has been applied successfully in many different domains [10], it has not seen the same level of attention as collaborative filtering techniques have. In recent years, competitions like the Netflix Prize<sup>1</sup>, CAMRA<sup>2</sup>, and the Yahoo! Music KDD Cup 2011 [6] have spurred on advances in collaborative filtering and how to utilize ratings and usage data from a community of users. However, there are many recommendation domains and applications where item content and metadata play a key role, either in addition to or instead of ratings and implicit usage data. For some domains, such as movies, the relationship between content and usage data has seen thorough investigation already (e.g. [12]), but for many other domains, such as books, news, scientific articles, and Web pages we still do not know *if* and *how* these data sources should be combined to provide the best recommendation performance. On top of that, in certain contexts of search, content-based recommendation plays a role as well. Personalised IR is strongly related to recommendation, but recent research in book search also identified search scenarios that combine aspects of retrieval and recommendation [7].

---

<sup>1</sup><http://www.netflixprize.com/>

<sup>2</sup><http://www.dai-labor.de/camra2010/>

---

The *CBRecSys* workshop series aims to address this by providing a venue for papers dedicated to all aspects of and new trends in content-based recommender systems. This includes both recommendation in domains where textual content is abundant (e.g., books, news, scientific articles, jobs, educational resources, and Web pages) as well as dedicated comparisons and combinations of content-based techniques with collaborative filtering approaches. After a successful first edition organized at RecSys 2014 in Silicon Valley [3, 4], the second edition of the workshop, CBRecSys 2015, was co-located with RecSys 2015 in Vienna, Austria.

## 2 Workshop Focus & Format

Authors were encouraged to submit papers on a variety of topics related to content-based recommendation, including:

- Developing novel recommendation approaches
- Exploiting user-generated content for recommendation
- Processing text reviews
- Mining contextual data from content
- Addressing limitations of recommender systems

In particular, papers submitted to the the workshop focused on the following topics. Several papers looked at how semantic information in the form of Linked Open Data resources can be integrated in recommendation algorithms. Other papers investigated solutions to common problems for recommender systems, such as ensuring diverse recommendations and dealing with continuous cold-start scenarios. Finally, a handful of submissions proposed novel ways of incorporating metadata features into collaborative filtering algorithms for domains such as fashion and books. A total of 12 papers were submitted, of which 8 were accepted for an oral presentation (for an acceptance rate of 66%). All papers were reviewed by at least 3 reviewers. In terms of attendance, CBRecSys 2015 proved to be a popular workshop with around 40 attendees.

### 2.1 Keynote

The workshop started with a keynote presentation on how to capture user interests for content-based recommender systems by Frank Hopfgartner (University of Glasgow, UK). This keynote gave a great introduction to the concepts and ideas behind content-based recommendation and the advantages of using semantic metadata. The presentation highlighted the challenges that come with an automated analysis of content, especially in the multimedia domain and identified different methods of capturing user interest. The keynote finished with a description of and invitation to participate in the NewsREEL challenge<sup>3</sup>, an evaluation campaign that allows researchers to benchmark news article recommendation algorithms in both an offline and online setting. The challenge offers a great platform for investigating content-based and hybrid recommendation techniques.

---

<sup>3</sup>Available at <http://www.clef-newsreel.org/>, last visited October 7, 2015.

---

## 2.2 Accepted Papers

This section briefly summarizes the eight accepted papers presented at the workshop. In their work, Tomeo et al. [14] addressed the issue of diversity in recommendation and proposed a model to infer intents from user profiles. User intents are modelled by multiple attributes of items using regression trees. The approach is evaluated on data sets from the domains of books and movies, with rule-based diversification leading to good diversity in recommended items across users, and in combination with traditional intent-aware algorithms leading also to good diversity within a single user's recommendations.

Musto et al. [11] presented a graph-based approach to recommendation by extracting item features from the Linked Open Data cloud. They compared a range of feature selection techniques such as PageRank, PCA, SVM and Information Gain, and found that specific selection techniques can be used to optimize for accuracy or diversity. The overall method outperforms several state-of-the-art baselines.

Kula [8] introduced LightFM, a hybrid model to address the cold-start problem for domains where item metadata is diverse and sparse. LightFM is a hybrid matrix factorization model that uses embeddings of metadata and natural language descriptions as latent features to represent both items and users, with which it can easily generalize to new items and users. The model naturally blends metadata and natural language descriptions into a single model.

De Graaff et al. [5] developed and evaluated a knowledge-based recommender system that is domain-independent, language-independent, and independent of the used social medium. The user interests are derived from the users social media profile, enriched with information from DBpedia. An user study showed that this very generic model can improve a domain specific baseline, indicating the complementarity of generic interests.

Bernardi et al. [2] argued that in e-commerce domains such as travel, recommender systems have to deal with continuous cold-start problems for known users, due to users returning infrequently, in different roles, with interests that change as users go through different phases of their life. The paper explores this problem and points out some important consequences, supported by examples from Booking.com, a major travel recommendation website.

Poussevin et al. [13] proposed to augment recommender systems by generating personalized reviews associated to items, as well as predict ratings and providing personalised sentiment classification. To personalize the reviews for a user, similar users are found based on the textual similarity between their reviews and a user's own reviews. After that, extractive summarization is used to generate the personal review.

Benkoussas and Bellot [1] presented a hybrid algorithm for solving the problem of book search and recommendation. They combined different retrieval models for the content-based recommendation part with a graph-based algorithm.

Labille et al. [9] proposed a recommender system for CiteSeerX, a digital library for computer science research, that takes both the content and the impact of articles and papers into account. Their hybrid recommender integrates scientific impact in the form of the authors' h-indexes as a part of the ranking criteria and was shown to produce more relevant paper than the baseline content-based recommender system already in use at CiteSeerX.

---

## 2.3 Breakout Session on Current and Future Challenges

The workshop was concluded by a breakout session on current and future challenges in content-based recommender systems. A research topic that came up throughout the session is the order in which to recommend items, the order in which they are consumed, and the order in which new items become available to recommend. This is a well-known issue in real-time scenarios in social media and news, but there are many other domains in which this plays an important role, such as scientific literature, health, job search (e.g. modeling career paths) and especially in e-learning. Issues such as novelty and tracking the development of a user's knowledge and interests and the development of topics and genres as new items become available is poorly understood.

Another important property of a good recommender system is the ability to generate explanations for why the items are being recommended and what kind of data is used to recommend them. Participants agreed that the relative importance of this depends on the application domain, with participants from different backgrounds arguing for the relative importance in their domain.

A recurring topic in recommender systems and IR research in general is diversity. Being able to produce diverse recommendations was seen as another important property of a good content-based recommender, although the necessity depends on the domain: for scientific articles or music, diversity is more important than for recommending job listings. Participants also expressed a desire for more research on whether and how users should be able to control the diversity and novelty of different recommendation algorithms.

The participants also discussed the desire for and problems with a RecSys challenge. Everyone agreed that a joint challenge in which many teams tackle the same problem and use the same data set and evaluation setup is a great way to advance the state-of-the-art in (content-based) recommendation. However, such joint challenges reveal a tension between the interests of academia and those of industry. Researchers in industry have their own, very specific problems to deal with and are understandably hesitant to invest time and resources elsewhere. Questions raised during the discussion were: What are topics of interest for both academics and industry? What are the criteria that such a challenge should satisfy for industry to consider spending their resources on it? What are domains where there are good resources for data and shared interest in research problems? It was suggested that Kaggle<sup>4</sup> is a good platform for getting many participants, but with the severe caveats that it does not allow challenge organizers to use their own evaluation measures and that there is little overlap between the communities of Kaggle and RecSys, with Kaggle participants feeling little urge to go to workshops and conferences to present their work. When asked about the possibility for contributing data to such a challenge, it was evident that companies are less likely to do so after the privacy issues of the AOL search logs and the first Netflix prize. Nevertheless, living lab initiatives for recommendation—such as the NewsREEL news recommendation challenge organized in conjunction with Plistawere seen as steps in the right direction.

---

<sup>4</sup>Kaggle is a platform for data mining and predictive modeling competitions; see <http://www.kaggle.com/> for more information.

---

### 3 Future Workshop

The continued popularity of the CBRecSys workshops in terms of both submissions and participation clearly shows a consistent interest in research on content-based recommendation. We are therefore planning a third edition of the CBRecSys workshop at RecSys 2016 in Boston.

### 4 Acknowledgments

We would like to thank ACM and RecSys for hosting this workshop, the RecSys workshop chairs Alan Said and Tsvi Kuflik. We would also like to thank the program committee: Robin Burke, Iván Cantador, Federica Cena, Paolo Cremonesi, Ernesto William De Luca, Tommaso Di Noia, Peter Dolog, Soude Fazeli, Cristina Gena, Juan Fernando Huete, Jaap Kamps, Birger Larsen, Babak Loni, Pasquale Lops, Cataldo Musto, Casper Petersen, Shaghayegh Sahebi, Alan Said, Giovanni Semeraro, Nafiseh Shahib, Marko Tkalčić, and Bei Yu. Final thanks are due to the paper authors, the keynote speaker Frank Hopfgartner, and the participants for a great and lively workshop.

The workshop material (list of accepted papers, invited talk, and the workshop schedule) can be found on the CBRecSys 2015 workshop website at <http://humanities.uva.nl/~mkoolen1/CBRecSys15/>. The proceedings were published as a CEUR Workshop Proceedings volume, available at <http://ceur-ws.org/Vol-1448/>.

### References

- [1] C. Benkoussas and P. Bellot. Cross-Document Search Engine For Book Recommendation. In *CBRecSys '15: Proceedings of the 2nd Workshop on New Trends on Content-Based Recommender*, pages 42–49, 2015.
- [2] L. Bernardi, J. Kamps, J. Kiseleva, and M. J. I. Müller. The Continuous Cold-start Problem in e-Commerce Recommender Systems. In *CBRecSys '15: Proceedings of the 2nd Workshop on New Trends on Content-Based Recommender*, pages 30–33, 2015.
- [3] T. Bogers, M. Koolen, and I. Cantador. Workshop on New Trends in Content-based Recommender Systems: (CBRecSys 2014). In *RecSys '14: Proceedings of the Eighth ACM Conference on Recommender Systems*, pages 379–380, 2014.
- [4] T. Bogers, M. Koolen, and I. Cantador. Report on RecSys 2014: Workshop on New Trends in Content-Based Recommender Systems. *SIGIR Forum*, 49(1):20–26, 2015.
- [5] V. de Graaff, A. van de Venis, M. van Keulen, and R. A. de By. Generic knowledge-based Analysis of Social Media for Recommendations. In *CBRecSys '15: Proceedings of the 2nd Workshop on New Trends on Content-Based Recommender*, pages 22–29, 2015.
- [6] G. Dror, N. Koenigstein, Y. Koren, and M. Weimer. The Yahoo! Music Dataset and KDD-Cup '11. In *JMLR Workshop and Conference Proceedings*, volume 18 of *Proceedings of KDD Cup 2011*, pages 3–18, 2012.

- 
- [7] M. Koolen, T. Bogers, A. van den Bosch, and J. Kamps. Looking for books in social media: An analysis of complex search requests. In *Advances in Information Retrieval: 37th European Conference on IR Research (ECIR 2015)*, LNCS. Springer, 2015.
- [8] M. Kula. Metadata Embeddings for User and Item Cold-start Recommendations. In *CBRecSys '15: Proceedings of the 2nd Workshop on New Trends on Content-Based Recommender*, pages 14–21, 2015.
- [9] K. Labille, S. Gauch, and A. S. Joseph. Conceptual Impact-Based Recommender System for CiteSeerx. In *CBRecSys '15: Proceedings of the 2nd Workshop on New Trends on Content-Based Recommender*, pages 50–53, 2015.
- [10] P. Lops, M. de Gemmis, and G. Semeraro. Content-based Recommender Systems: State of the Art and Trends. In *Recommender Systems Handbook*, pages 73–105. 2011.
- [11] C. Musto, P. Basile, M. de Gemmis, P. Lops, G. Semeraro, and S. Rutigliano. Automatic Selection of Linked Open Data Features in Graph-based Recommender Systems. In *CBRecSys '15: Proceedings of the 2nd Workshop on New Trends on Content-Based Recommender*, pages 10–13, 2015.
- [12] I. Pilászy and D. Tikk. Recommending New Movies: Even a Few Ratings Are More Valuable Than Metadata. In *RecSys '09: Proceedings of the Third ACM Conference on Recommender Systems*, pages 93–100. ACM, 2009.
- [13] M. Poussevin, V. Guigue, and P. Gallinari. Extended Recommendation Framework: Generating the Text of a User Review as a Personalized Summary. In *CBRecSys '15: Proceedings of the 2nd Workshop on New Trends on Content-Based Recommender*, pages 34–41, 2015.
- [14] P. Tomeo, T. D. Noia, M. de Gemmis, P. Lops, G. Semeraro, and E. D. Sciascio. Exploiting Regression Trees as User Models for Intent-Aware Multi-attribute Diversity. In *CBRecSys '15: Proceedings of the 2nd Workshop on New Trends on Content-Based Recommender*, pages 2–9, 2015.