8 Conclusions

This thesis describes large scale real time on line video recommendation system. Currently in use by the Internet branch of Globo Group, the leader in the broadcasting media segment for Brazilian Internet audiences, this thesis demonstrated an innovative design that marries state-of-art recommendation techniques to emerging cloud computing technology.

In particular, it argued that the combination of mature research in collaborative filtering for item-item recommendations and emergent cloud computing technology opens up a great number of research challenges and possibilities. It introduced an architecture for real-time item-item recommendations, which rationalizes the use of resources by exploring on-demand computing.

This thesis also presented an adaptive implicit feedback model to be used in a recommendation system based on collaborative filtering, which goal is to obtain a more efficient interest inference than the implicit binary feedback. The first step was to understand what are the characteristics of video consumption that impacts in the efficiency of the implicit binary feedback inference of interest, for example in short breaking news and sports videos. Based on consumption patterns, it proposed a feedback model able to adapt efficiently to different consumption patterns, and dynamically adapts itself to the variations of these patterns. The model is able to be individualized by the video, which leverages to a more accurate inference of interest for a particular content.

To validate the feasibility of the proposed collaborative filtering architecture and implicit feedback model, a real case of implementation of this new approach in a production environment in Internet Industry, with experimental results that demonstrate the feasibility of the proposed approach for environments with dozens of millions of users and items. For Globo.com, this real-time recommendation architecture increased the end screen video suggestion conversion rate (number of users that clicks in a recommendation by the total times that the recommendation is shown) from 14%, using a traditional offline collaborative filtering approach, up to

25%, which directly impacted business metrics such as ad impressions, unique visitors and visitor retention, at an almost insignificant cost when compared with all direct and indirect benefits. This real-time recommendation architecture is still being used by Globo.com in their video platform and was also tested for articles recommendations in their news portal, which also presented a better conversion rate then the offline similarity calculation approach, according to the first collected statistics.

The fact that the elastic capacity of a public Cloud such as the AWS is very large, i.e., it offers access to practically unlimited resources, changes the fundamental premises that hitherto guided the research in this area and has the potential implications given below:

- The (recommendation quality x processing time) trade-off, no longer holds. New architectures can be developed aimed at maximum recommendation quality without input data sampling or model simplification, as processing times can be reduced by using parallel implementations such as the proposed approach.
- Choosing an optimal recommendation algorithm is known to be very difficult. The possibility of using several approaches, in parallel and at a low cost, may herald the beginning of a new era of experimentation in content suggestion.
- Adapting parallel approaches to work with linear algorithms requires a great deal of effort. In the case of the proposed approach, it was developed a multi-layer architecture to allow for the use of independent scaling processes for each layer by addressing each processing layer independently. Although the results are encouraging, much remains to be done. In particular, it is possible that the development of self-tuning algorithms to determine optimal instance management could achieve very good results.

Finally, it is important to highlight that the techniques used for the similarity calculations are versatile, i.e., they can be exchanged and customized as needed. This ensures flexibility, adaptation, extensibility, and generality of the proposed

architecture. This result also shows that the use of the proposed model for inference of interest is promising for video recommendations.

As future work, it is important to make a more detailed and focused evaluation of this model, using scales with more and fewer levels, analyzing conversion segmented by category, duration, etc. This kind of evaluation could highlight what is the impact of different levels in the recommendation efficiency.

Other opportunity is to evaluate the proposed architecture with different types of content besides video. Since it is agnostic of the content and its characteristics, it could also be adapted to be used for products, pages, songs, etc. The only adaptation needed is regarding how the implicit feedback should be acquired. For example, for pages it could be the time spent on it, or, for songs, something similar to video, based on playback duration, etc.

It could be also interesting to change the similarity metric to see the impact in the recommendation efficiency. The proposed architecture strongly relies on the cosine similarity. Change it to work with a different similarity metric could have an important impact in the conversion rate.

Finally, the proposed architecture introduces a new similarity metric for multi-valued feedbacks that were decomposed in binary feedbacks, through the consolidation of cosine similarity. Find different ways to decompose and consolidate the multi-valued feedbacks could also have a strong impact in the recommendation efficiency.