Applying Recommendation Algorithms to Collections of Geospatial Metadata

P. Mooney[1, 2] and A.C. Winstanley[1]

[1]National Center for Geocomputation, National University of Ireland Maynooth, Co. Kildare, Ireland
Telephone: +353 1 708 6455
Fax: +353 1 708 6455
Email: peter.mooney@nuim.ie, adam.winstanley@nuim.ie

[2]Environmental Research Center, Environmental Protection Agency, Richview, Clonskeagh, Dublin 14. Ireland
Telephone: +353 1 268 0100
Fax: +353 1 268 0199
Email: peter.mooney@nuim.ie

1. Introduction

In Spring 2007 the Environmental Research Center (ERC) of the Irish Environmental Protection Agency (EPA) launched the SAFER system – Secure Archive For Environmental Research Data. SAFER is a fully web-based system (available at http://coe.epa.ie/safer) for the collection, maintenance, display, and long-term archival of metadata and raw data from EPA funded environmental research projects in Ireland. All project managers are compulsory required to provide metadata and the raw data generated within their projects. SAFER maintains all metadata within a MySQL database and adheres to the ISO 19115 metadata standard. As in Lagoze, (2001) the metadata is stored as a single information object but is projected to multiple different views depending on the audience and context. Raw data is stored in a Storage Area Network.

During the period 2007 – 2013 the EPA will fund over €100 million of environmental research in Ireland. Over this period it is anticipated that a very large volume of raw and aggregated geospatial data will be generated and collected. Several hundred fully ISO 19115 compliant metadata records will also be collected. While this will be a very successful utilisation of the SAFER system the quantity of metadata resources may give rise to several problems. This abstract outlines research into implementing Collaborative Filtering (CF) algorithms into the metadata search and browsing components of SAFER. These algorithms would make recommendations about other related metadata resources in SAFER to users based on the preferences and usage patterns of other similar users. The remainder of this abstract outlines a brief example of the weakness of traditional search and browse techniques in large metadata collections. We then briefly outline how the CF algorithms are being implemented and a summary of their expected impact and results.

2. Searching and Browsing Metadata Collections

SAFER uses metadata to provide research scientists, environmental policy makers, and concerned citizens with a means to discovering, downloading, and using environmental research data collected from funded programmes. Some metadata attributes draw their values from controlled vocabularies (CV). Other attributes uses a free text keyword (FT) system (referred to by Hossain et al (2006) as a *folksonomy*). The entire taxonomy of metadata on SAFER can be searched and browsed using the standard classification attributes:

- *ISO19115 Topic Category* (19 Topic Categories for Geospatial Data) CV
- *EPA project category* (8 Project Categories – specified internally by EPA) CV
- *Keyword search* (Pattern-based matching on collection of keywords) FT
- *Temporal search* (searching constrained by the temporal range of the resources)
- *Responsible party search*. (Pattern-based matching of owner information) FT

Within environmental science users' interest are seldom restricted to a specific topic or category. There is often a large degree of thematic overlap. Take the following example. A researcher views a metadata resource (A) related to a land-cover analysis project (category=land). There exists a metadata resource (B) describing a project on phosphorus run-off analysis (category=soils). In a very large collection of metadata this researcher may fail to find the metadata resource B. In this context of this example the "search mechanism is precise (finding A) yet brittle and the browsing facility is robust buy vague (possibly missing B)" (Rao, 2003). A similar situation is imagined where a Strategic Environmental Assessment (SEA) practitioner is searching for data resources by category when search based on geographical location is more suitable. The spatial locational context of the related metadata resources is often not included in search implementations.

Providing users with the facility to search and browse the metadata collection is not difficult from a software implementation point of view. However, we feel that this functionality needs to extend beyond traditional sorted listings of matching search results. The potential of using the *data contained within the* metadata (geospatial, temporal) and *data attached to the metadata* (temporal information of when users accessed it, viewed it, downloaded it, etc) should be explored. This point is emphasised by Bulterman (2004) where the author argues that "one of the great paradoxes of our time is that although more information is being searched for than ever, the correct use of conventional metadata is probably at a 10 year low".

3. Application of Collaborative Filtering Algorithms to Geospatial Metadata

Applying CF algorithms to SAFER's metadata collection is based on the idea that users browsing the metadata and associated geospatial data resources should be able to take advantage of what other users have already browsed and evaluated. We assume that the scientific relationships between metadata resources will remain relatively static over time. As the number of metadata resources in a metadata collection such as SAFER increases the difficulty for users to discover related resources also increases. The collaborative filtering algorithm we are implementing is that described in Linden et al (2003). The algorithm is usually computed offline and has a run-time complexity of $O(N^2M)$ where N is the number of items (metadata resources) and M the number of registered users. This item-centric approach will attempt to find releationships between metadata items and make recommendations based only upon user preferences and these relationships. This CF algorithm is implemented at Amazon.com recommending similar products based on a user's specific actions in viewing, browsing, rating, and buying other products.

Passive filtering data collection is the method by which SAFER collects user browsing and preference information. We only analyse information from registered system users (when they are logged on). As users navigate through SAFER their navigation paths and choices are recorded. This implicit filtering relies on the actions of the user to determine a value rating for specific content. Actions used by the scoring mechanism include: downloading datasets; repeatedly viewing or printing a metadata record; types of search terms; viewing content on external links or maps. The temporal information collected is a important characteristic of the collected information. The logged timestamp of when a user viewed a metadata record and performed their next action can be used to determine where they just scanned the metadata information or where it could be reasonably assumed that performed a more detailed examination of it. This implicit data is easy to collect in large quantities and no extra efforts are required on part of the user. It is stored within the SAFER database and is easily linked to metadata resource tables and user profile tables.

There are several important algorithmic considerations. Amazon.com run the algorithm mentioned above offline as their N and M values are in the order of millions. We execute the algorithm on 6 hourly intervals offline and automatically update the recommendation database table. The algorithm must exhibit good *cold start ability:* that is it should provide good recommendations to a user who is new to the system and who has not performed any actions as yet. The cosine measure of the angle between two vectors is used to calculate the similarity between metadata resources (Linden et al, 2003; Kangas, 2002). We are also investigating how including the actions of non authenticated users in the scoring mechanism improve/skew the recommendations.

4. Concluding Remarks

As the quantity of geospatial information continues to grow at exponential rates the requirement to properly document this information becomes ever more crucial. In addition to this there is an exponential increase in the number of users with little or no geospatial data manipulation experience using geospatial data services (Doughty et al, 1999). Metadata recommendations for users of geospatial metadata are computed by finding metadata resources that are similar to other resources the user has viewed, accessed, or downloaded. This provides users with added-value browse and search functionality on large collections of metadata resources. CF algorithms have many possibilities yet to be exploited on the Internet. Presently it is often used as a "light-weight alternative to more data intensive processes such as data mining using neural networks or rule-based learning" (Kangas, 2002). The computational effectiveness of metadata will always be somewhat dependent on the human expert providing *good* metadata. The trade-off between the costs of developing and managing metadata collections against the types of search functional and potential future uses must always be seriously considered.

Acknowledgements

This work is being carried out as a Postdoctoral Fellowship in the Environmental Protection Agency of Ireland (EPA) within the Environmental Research Center (ERC). The project is funded under the National Development Plan 2000 – 2006.

References

Doughty, J., Jones, P., Nolan, J., and Hirsh, S. 1999, Interoperable Geospatial Objects Proceedings of GeoComputation 99, Mary Washington College, Fredericksburg, VA, USA. 25-28th July 1999.

Bulterman, D.C.A. Is it time for a moratorium on metadata? IEEE Transactions on Multimedia, 11 (4) pp 10—17, 2004.

Hossain, M.A., Rahman, M.A., Kiringa, I., and El-Saddik, A. Metadata Tagging and Mapping Framework for Managing Multimedia Content. Proceedings of the 8th IEEE International Symposium on Multimedia. 2006

Kangas, S. Collaborative Filtering and Recommendation Systems. Research Report Number TTE4-2001-35, VVT Information Technology, P.O. Box 12041, FIN-02044 VTT, Finland. Jan 17th 2002. http://www.vtt.fi/tte (last accessed April 2007)

Lagoze, C. Keeping Dublin Core Simple: Cross Domain Discovery or Resource Description? D-Lib Magazine, 7(1) January 2001. http://dlib.anu.edu.au/dlib/january01/lagoze/01lagoze.html (accessed April 2007)

Linden, G., and Smith, B., and York, J. Amazon.com Recommendations Item-To-Item Collaborative Filtering. IEEE Internet Computing Pages 76 – 80. January 2003.

Rao, R. From Unstructured Data to Actionable Intelligence. IEEE Magazine on Knowledge Management IT Pro. Pages 29 – 35. November-December 2003