The Curated Web: A Recommendation Challenge

Zurina Saaya[1], Rachael Rafter[2], Markus Schaal[1] and Barry Smyth[2]
[1]CLARITY: Centre for Sensor Web Technologies, University College Dublin, Ireland
[2]INSIGHT: Centre for Data Analytics, University College Dublin, Ireland
{firstname.lastname}@ucd.ie

ABSTRACT

In this paper we consider the application of content-based recommendation techniques to web curation services which allow users to curate and share topical collections of content (e.g. images, news, web pages etc.). Curation services like Pinterest are now a mainstay of the modern web and present a range of interesting recommendation challenges. In this paper we consider the task of recommending collections to users and evaluate a range of different content-based techniques across a variety of content signals. We present the results of a large-scale evaluation using data from the Scoop.it[1] web page curation service.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Filtering; H.3.5 [**Online Information Services**]: Web-based Services

Keywords

content-based recommendation, social web, curation

1. INTRODUCTION

Recently there has been a shift in web usage from content consumption to content production and a pattern of *content curation* has emerged through sites and services like Pinterest, Tumblr, Storify, and Scoop.it. These curation sites allow users to curate collections of content (e.g. images, news, web pages etc.) on topics that matter to them and to share these collections with others.

In this paper we consider some of the recommendation challenges that exist within such curation services using Scoop.it (a web page curation site) as an evaluation case-study. Specifically, we consider the challenge of matching users with collections. We look at different ways to profile the interests of users and the contents of collections in order to develop a content-based recommendation framework to suggest relevant collections to users based on their current collections/topics/interests. We evaluate the efficacy of different types of content signals during user profiling and collection indexing, from high-level collection descriptions to detailed information about the contents of individual pages within a collection. We do this to better understand the relationship between these different types of signals and recommendation performance, and describe the results of a large-scale evaluation based on Scoop.it user and collection data.

Social curation is a relatively new trend and has received little attention to date. In previous work we examined how to assist curators during the content curation phase, using machine learning to identify the correct target collection for new content the user wishes to gather [7]. In [5] the authors look at recommending curators, identifying the important users that form the community around a news story on Twitter using network analysis. Others focus on how users curate tweets, and how to recommend new tweets to them [4]. The work in [2] also studies content recommendation on Twitter, using both content and social signals.

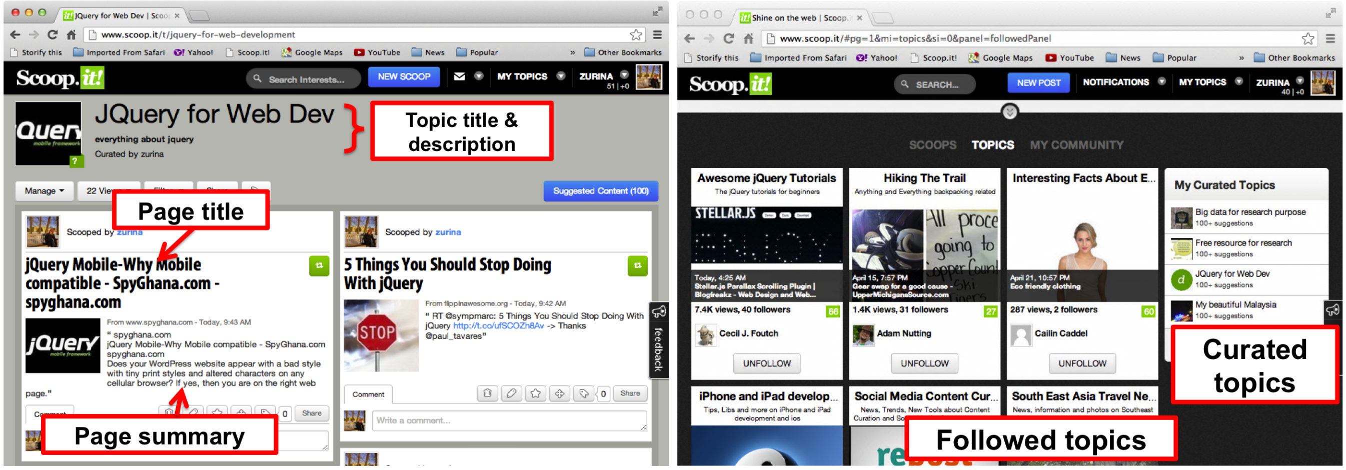
2. WEB CURATION USING SCOOP.IT

Scoop.it is a web-based curation tool that provides a platform for users to curate all of their favorite resources on a given topic for sharing with interested parties. Users can create their own collections (or *topics* in the Scoop.it parlance) and the service provides a range of tools to help users identify and filter content for their collections. Creating a collection or topic is a simple matter of providing a name, a short description, and a set of suitable keywords.

For instance, a curator may create a new *JQuery for Web Dev* topic (or collection) as a place to curate all things related to web development using *jquery* (see Fig. 1(a)). For each topic, the user can specify a title and short description of the topic, and every page added to the topic is associated with a page title and short summary. Furthermore, curators can annotate the pages that they include in their topic with their insights, and others can comment on the pages included. In addition to curating their own topics, users can also follow the topics of others that interest them, (see Fig. 1(b)). This can be a useful way for users to maintain and improve their own topics, by keeping up-to-date with what others are curating in similar topics. Scoop.it provides a handy *rescoop* functionality, allowing users to conveniently incorporate content from other topics into their own.

[1] http://www.scoop.it

RecSys'13, October 12–16, 2013, Hong Kong, China.
Copyright 2013 ACM 978-1-4503-2409-0/13/10 ...$15.00.
http://dx.doi.org/10.1145/2507157.2507216.



(a) Topic page (b) Topics dashboard

Figure 1: Scoop.it in action

3. RECOMMENDER FRAMEWORK

We describe a content-based recommendation framework that explores the relative benefits of different sources of data for modelling curated collections and user profiles.

3.1 Representing Topics for Querying and Indexing

In Scoop.it each topic is a collection of curated pages and has an associated topic title and description. Each page in a topic has a URL, a title, and a summary (snippet) description. In addition we can also extract a set of content terms from the page directly. This means that we can represent and index topics at different levels of granularity:

1. *TopicTitleDesc* - The terms in the topic's title and description elements.
2. *PageTitleSumm* -The combination of terms from the page titles and summary descriptions of all pages in the topic.
3. *PageContent* - The combination of the top-N terms taken from the full content of each page in the topic.

In addition to representing topics with raw terms as above we also produce *feature-based* descriptions by applying LDA [1] and LSI [3] to both page-level and content-level terms to produce *LSIPageTitleSumm*, *LSIPageContent*, *LDAPageTitleSumm*, *LDAPageContent* representations.

This produces 7 different content *types* that can be used as the basis of collection indexing and user profiling. In the case of the former we use a standard indexing approach in which each topic is represented as a document vector based on the different content types. So in the case of *TopicTitleDesc* each topic (t_i) is represented as a vector of the topic's title and description terms; we refer to this as $I(t_i, TopicTitleDesc)$.

Users in Scoop.it are associated with the topics they curate (create) and other topics they choose to follow. Therefore we can profile a user u_i based on her curated topics ct^{u_i}, or her followed topics ft^{u_i}, or both $ctft^{u_i}$; we refer to these as the *source* of the profile data. And then we can represent each of these profile types using one of the 7 different approaches above. In other words, if we profile u_j from her curated topics using *LSIPageContent* then the profile will be made up of the LSI features extracted from the page content terms of all pages in the user's curated topics only; we refer to this as $P_{ct}(u_j, LSIPageContent)$.

3.2 Recommending New Topics

We adopt a simple retrieval-based content recommendation approach, treating user profiles as queries against the topic index to recommend the n most similar topics. In fact, we take a weighted retrieval approach using Lucene's TF-IDF weighting metric during retrieval [6]; for a given user u_j, we score a topic t_i based on those terms/features it shares with the user's profile $P = P_{src}(u_j, type_u)$, in proportion to their frequency in u_j's profile and inversely proportional to their frequency in the topic index $I = \bigcup_{t_i} I(t_i, type_I)$ as a whole.

$$Score(u_j, t_i, src, type_u, type_I) = \sum_{e \in t_i \cap u_j} tf(e, P) \times idf(e, I)$$

Once again it is worth stressing that our aim here is not to propose a novel recommendation technique but rather establish an approach that facilitates a *like-for-like* comparison between different types of profile/index representation and profile source. The above combinations of profile representation and source, and indexing data, accommodate many different combinations of recommendation techniques: 3 sources of profiling data, 7 types of profile, and 7 types of index for a total of 147 recommendation configurations.

4. EVALUATION

The purpose of this evaluation is to consider how the different types of indexing data and profiling sources impact overall recommendation quality when it comes to suggesting new topics to users. To do this we ran an offline, training-test style recommendation study over Scoop.it data scraped from their API during October-November 2012. This data include 22,000 unique topics covering more than 2 million

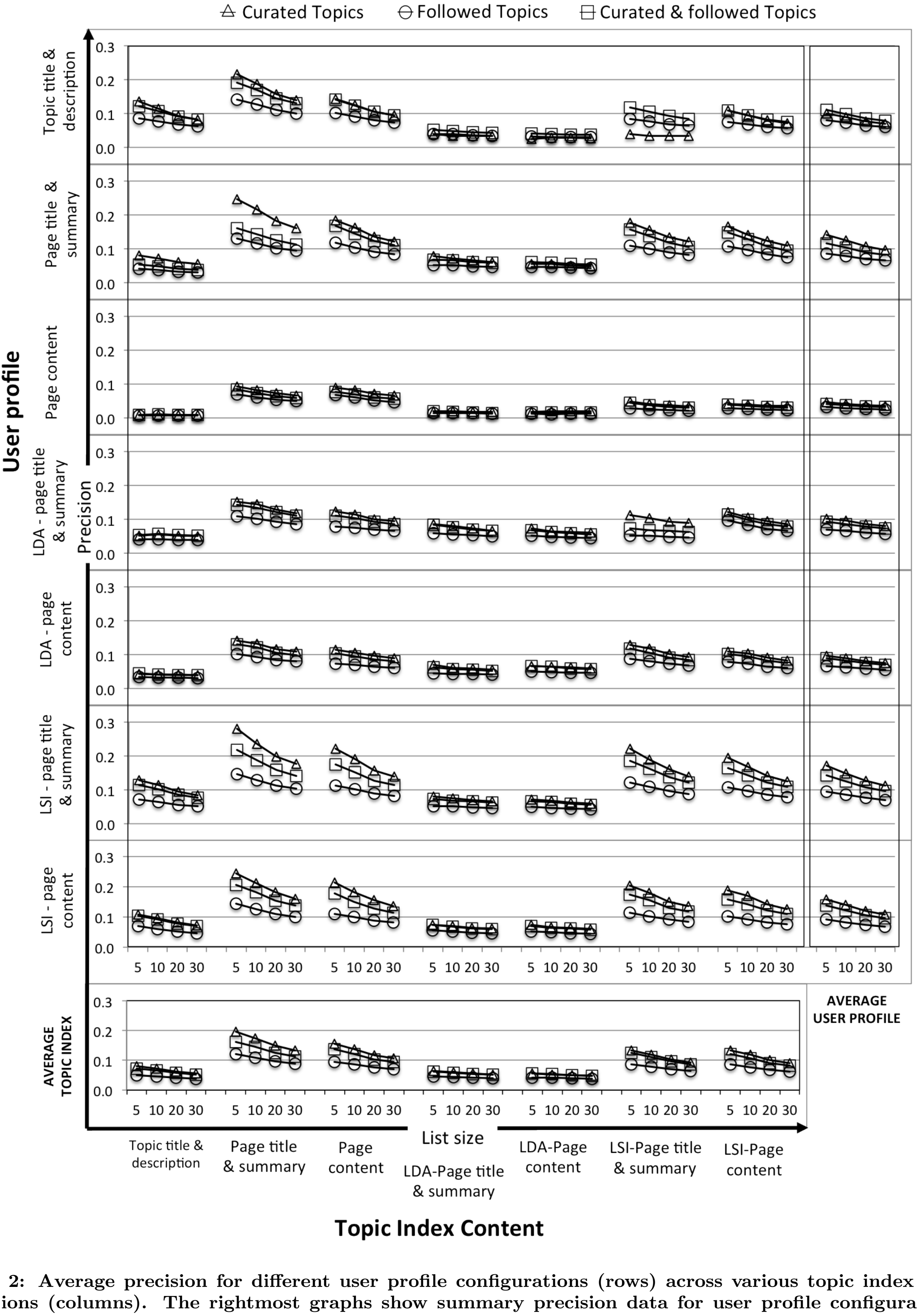


Figure 2: Average precision for different user profile configurations (rows) across various topic index configurations (columns). The rightmost graphs show summary precision data for user profile configurations (averaged across topic index configurations); similarly, graphs in the bottom row show summary precision data for topic index configurations (averaged across profile configurations).

pages and we focus on a subset of 845 active users who follow at least 20 topics; on average these users follow 90 topics and have created 5 topics of their own, each of which contains at least 100 pages. In this study the *training* instances are made up of user profiles containing all of the user's created topics plus random selections of 10% of their followed topics. The remaining followed topics serve as *test* instances. We do this 10 times to select different sets of followed topics for training and average our results across these folds.

For each set of training-test data we create all possible configuration combinations (7 x 7 x 3) of profile and recommendation index as described above. For each user a set of n (n = 5, 10, 20, 30) recommendations is made, and compared to the *followed* topics in the test profile in order to compute the average *precision*. Of course there are other standard measures we could apply here such as *recall*, *F1-measure* or *normalised discounted cumulative gain (NDCG)*; due to limited space however, we only consider precision at this point. The results are presented in Figure 2; the rows represent different types of profile data and the columns different types of index data. There are actually 3 sets of graphs presented in this figure: (1) the main *core* of 7 x 7 graphs shows the individual results of each of the (7 x 7 x 3) configurations that represent our recommendation design space; (2) the bottom row of 7 graphs (*average topic index*) represent summary precision data, for a given index configuration, averaged across the profile configurations (the average of the columns); and (3) the rightmost column (*average user profile*) shows a similar averaging across the rows, each profile configuration averaged across the index configurations.

There is a lot of result data presented but a number of interesting patterns can be observed. Focusing on the core graphs we can see that there is considerable precision variation across the different profiling and indexing configurations; precision results vary from <10% to almost 30%. And so the choice of profiling and indexing data matters. Generally speaking better results are observed when using curated topics versus followed or curated plus followed topics. Even though a given profile is only made up of a small fraction of curated topics this information provides a much clearer recommendation signal. For example, the results when using *page title* and *summary* information for profiling show a clear benefit for curated topics with precision ranging from about 0.25 to 0.15 (as n increases) compared to 0.2 - 0.1 for the followed and curated plus followed source configuration. Interestingly, this result aligns with the findings in [2] which show that the content in a user's own tweets (on Twitter) are a better signal of her information seeking preferences, than the content in the tweets of the people she follows.

The row and column averages help us to draw some general conclusions. We can see that extracting LDA features rarely helps, at least not as much as using LSI. And generally speaking basing recommendations on topic-level information from the *topic title* and *description* produces limited precision results. At the same time, using detailed *page content* (i.e. extracting actual term-content from pages) performs similarly poorly when used as the profile query but performs well when used to index topics; compare the *page content* configurations in the *average user profile* summary column to the *average topic index* summary row. Generally speaking it appears that page-level content, in the form of *page title & summary* configurations (using terms and LSI derived features) provides the right level of representational clarity to deliver the best performing recommendations. Specifically building a term-based index from *page title & summary* data and using LSI to extract *page title & summary* profiles gives the best overall recommendation results.

5. DISCUSSION

Our evaluation using Scoop.it data covered a comprehensive design space of recommendation configurations from which some general conclusions can be drawn. First and foremost, we demonstrated that relevant recommendations can be delivered even when evaluated against a traditionally conservative precision metric that focuses on those topics/collections only already followed by users. We often achieve precision scores up to 0.3 indicating that 30% of the recommended topics were found to be relevant in the sense that the user already followed them. Generally speaking, curated topics (those topics the user has created themselves) provided a strong recommendation signal than those they followed and delivered better precision, despite their relative sparsity compared to followed topics. And finally, the representational granularity offered by the intermediate *page title & summary* data provided superior precision results, although indexing (but not profiling) using detailed *page content* information also performed well.

Acknowledgments

This work is supported by Science Foundation Ireland under grant number 07/CE/I1147, Ministry of Education Malaysia and Universiti Teknikal Malaysia Melaka. The INSIGHT Centre for Data Analytics is supported by Science Foundation Ireland under Grant Number SFI/12/RC/2289.

6. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[2] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: experiments on recommending content from information streams. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 1185–1194. ACM, 2010.

[3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[4] K. Duh, T. Hirao, A. Kimura, K. Ishiguro, T. Iwata, and C. Yeung. Creating stories: Social curation of twitter messages. In *Sixth International AAAI Conference on Weblogs and Social Media*, pages 447–450, 2012.

[5] D. Greene, F. Reid, G. Sheridan, and P. Cunningham. Supporting the curation of twitter user lists. In *NIPS 2011 Workshop on Computational Social Science and the Wisdom of Crowds*, Sierra Nevada, Spain, December 2011.

[6] E. Hatcher and O. Gospodnetic. *Lucene in action.* Manning Publications, 2004.

[7] Z. Saaya, M. Schaal, R. Rafter, and B. Smyth. Recommending topics for web curation. In *User Modeling, Adaptation, and Personalization*, pages 242–253. Springer, 2013.