Serendipitous Browsing: Stumbling through Wikipedia

Claudia Hauff and Geert-Jan Houben
Web Information Systems
Delft University of Technology
Delft, the Netherlands
{c.hauff,g.j.p.m.houben}@tudelft.nl

ABSTRACT

While in the early years of the Web, searching for information and keeping in touch used to be the two main reasons for 'going online', today we turn to the Web in many different situations, including when we look for entertainment to pass the time or relax. A popular tool to facilitate the users' desire for entertainment is StumbleUpon, which allows users to "stumble" through the Web one (semi-random) page at a time. Interestingly to us, many StumbleUpon users appreciate being served Wikipedia articles, which are informative pieces of text that educate the reader about a particular concept. The leisure activity of stumbling can thus also incorporate a learning experience. Since life-long learning is an important characteristic of knowledge economies, it is crucial to understand the interplay between these two - at first sight - opposing forces. We hypothesize that a greater understanding of what makes certain Wikipedia articles more attractive to the serendipitously browsing user than others, will enable us to develop adaptations that expose a greater amount of Wikipedia articles to the leisure seeking user.

Categories and Subject Descriptors: H.3.3 Information Storage and Retrieval: Information Search and Retrieval
General Terms: Human Factors, Experimentation
Keywords: free-choice learning, educational leisure, serendipitous browsing

1. INTRODUCTION

In the early years of the Web, searching for information and keeping in touch used to be the two main reasons for 'going online'. Today, we rely on the Web in increasingly diverse situations including shopping, consultations and learning. While these examples are all directed towards a particular goal the user has, we also turn to the Web at times when we simply want to be entertained to pass the time or relax. The possibilities for entertaining yourself on the Web are manifold, one can play games, listen to music, watch movies or simply browse through the Web in the hope of finding entertaining pages. Due to the sheer size of the Web though, random browsing is not effective for discovering pages that may b interesting to the individual user. For this reason, a number of services have become popular that recommend web pages to users based on their interests. One popular tool to facilitate the users' desire for entertainment by serendipitous browsing is StumbleUpon[1] (SU), which allows users to "stumble" through the Web one (semi-random) page at a time. Interestingly to us, many SU users appreciate being shown Wikipedia[2] articles, which are informative pieces of text that educate the reader about a particular concept. The leisure activity of stumbling thus can also incorporate a learning experience, which might contribute to the development of novel ideas and lead to creative insights. Since life-long learning is an important characteristic of knowledge economies, it is crucial to understand the interplay between these two seemingly opposing forces (entertainment vs. learning). We hypothesize that a greater understanding of what makes certain Wikipedia articles more attractive to the serendipitously browsing user than others, will enable us to develop adaptations that expose a greater amount of Wikipedia articles to the leisure seeking user.

In this position paper we make an argument for the importance of this task. We draw from a number of insights gained in museum studies [11] where the question of how learning can be facilitated in leisure settings (the museum visit) has been investigated for many years. While we do not consider the SU pages to be similar to museum objects, we do find a number of parallels.

A first experiment on the stumbled Wikipedia pages revealed that, just as in museums not all objects are equally attractive to visitors, not all articles are interesting to the average StumbleUpon user. In fact, only a very small number of Wikipedia articles gather a large number of views by SU users, most articles are rarely viewed. While we have no answer yet to the question of how to automatically classify articles according to their attractiveness to the serendipitously browsing user, we have developed a number of hypotheses which are outlined in Section 3.2.

If we assume for a moment that we are indeed able to develop such an approach, a number of application scenarios can be envisioned:

- A qualitative study of the features that play a role in to trickling the interest of users who do not have an information need, will enable Wikipedia contributors to write their articles in a way that is more accessible to such users.
- Wikipedia is available in many different languages and such a prediction method would allow us to bootstrap a recommender like StumbleUpon in different languages by adding an initial set of interesting, high quality pages before the critical mass of users is reached.

Presented at Searching4Fun workshop at ECIR2012. Copyright ©2012 for the individual papers by the papers' authors. Copying permitted only for private and academic purposes. This volume is published and copyrighted by its editors.

[1] http://www.stumbleupon.com/

[2] http://www.wikipedia.org/

- Outliers (articles with many 'Likes' but a low probability of being attractive) can be manually investigated to reduce spam. Or conversely, undiscovered articles are obtained and can be injected into the index.
- The passages that trigger the surprise or the attractiveness of an article can be identified and highlighted to the browsing user. This may help to keep those serendipitously browsing users engaged that initially only quickly scan the article.
- E-learning applications can also benefit, as articles which are interesting to the casual reader can be found this way.

The rest of the paper is organized as follows: related work is presented in Section 2, followed by a preliminary analysis of stumbled Wikipedia pages (Section 3) and the conclusiosn (Section 4).

2. RELATED WORK

For this work, we draw inspirations from two areas. On the one hand we consider research into so-called *educational leisure settings* and *free-choice learning* which is a multidisciplinary field that includes aspects from sociology, psychology and education. On the other hand, our work is also strongly related to serendipity.

Education leisure settings can be found in a wide range of institutions including museums [12], national parks, zoos, science centers [5], etc. As the name suggests, these institutions serve two purposes: to educate the public as well as to provide an entertaining experience to the visitors. Education leisure settings can be characterized by a number of commonalities with respect to the visitors and their learning experience [9, 10, 11]: (i) the visitors gain direct experience, (ii) they decide what and whether at all to learn, (iii) the learning process is guided by their interests, (iv) learning is influenced by the visitors' social interactions and (iv) the visitors are a highly diverse group, with different educational backgrounds and prior knowledge. Since learning in this setting is voluntary, the visitors' motivation plays an important role: why did they come?

Serendipity, the act of encountering information nuggets unexpectedly, has mostly been investigated in the context of education [3] and work-related discoveries after serendipitous moments. One of the works outside of this realm is [6] where tools were developed to help people reminisce in their own digital collections. In goal-directed Web search the potential for serendipitous encounters has also been recently investigated [2], while [1] offers an insightful discussion of serendipity and how it is used, exploited and induced in computer science.

Finally we note that different aspects of Wikipedia articles have also been investigated in the past, though not from a perspective of serendipitously browsing users. For instance, in [7] it was found that the writing style distinguishes so-called featured articles in Wikipedia[3] from unfeatured articles. Classifying Wikipedia articles according to their quality, as defined by Wikipedia contributors, was also investigated in [13], where network motifs and graph patterns in the editor-article graph were exploited.

3. STUMBLEUPON

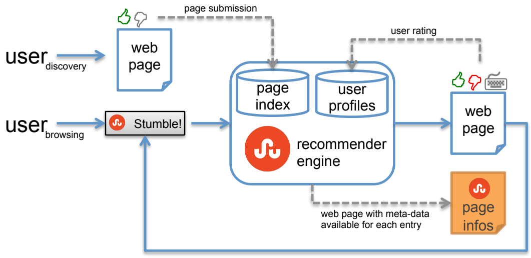

Figure 1: A StumbleUpon user can contribute Web pages he likes to the index and he can "stumble" pages that are in the SU index according to his interests. One page at a time is shown; the user can provide feedback in terms of like and dislike.

The usage of StumbleUpon is depicted in Figure 1. A user "stumbles" pages with a simple click of the 'Stumble!' button in his browser toolbar. In response, the user is presented with a random page from the Web, biased according to his user profile or his friends' 'Likes'. The simplicity of the system protects the user from information overload [8, 4], a user has only two choices when faced with a stumbled page: either to start reading or to continue stumbling. Users can also contribute pages to the SU index: whenever a SU user discover a web page that is not yet in the index and that he likes, he can add it by means of the 'Like' button. Finally, for each page in the SU index, there is a SU page which contains meta-data, including the number of users who viewed/liked the page, the category the user who discovered the page placed it in and the comments users left about the page.

3.1 Wikipedia Articles in StumbleUpon

In all experiments we report here, we utilize the English Wikipedia dump *enwiki-20111007* from October 2011. In a pre-processing step, we selected all Wikipedia articles that are neither redirects to other articles, nor new articles or explicit disambiguation pages and have a length of at least 500 characters (to remove stubs). In total, $3,552,059$ articles remained.

In order to determine the popularity of Wikipedia articles in StumbleUpon, we randomly selected half of these Wikipedia articles and queried the StumbleUpon API for their number of views by SU users. Since SU is a recommendation engine, we can safely assume that the highly viewed pages are also highly popular and liked. We note, that the number of 'Likes' a page has received is not accessible through the StumbleUpon API. The information is accessible though at the SU meta-data page, which we manually checked for the results reported in Table 1.

Among the evaluated $1,776,029$ articles, we found $267,958$ (15.13%) of them to be contained in the SU index. In our initial investigation, we also considered French and German Wikipedia which are two of the largest non-English Wikipedia repositories. However, we only found a very limited number of their articles in the SU index (in both cases less than 1%) and thus did not consider them further. Thus, an application scenario as proposed in the introduction (to bootstrap a recommender for a new language) is highly desirable.

Let us now focus on those articles that were submitted

[3] Featured Wikipedia articles are of particularly high quality and chosen by Wikipedia editors.

by Stumblers to the index. Figure 2 shows a scatter plot of the number of views versus the number of Wikipedia articles in the index. As can be expected, most articles have very few views (the median number of views is 10) while a small number of articles have gathered more than half a million views.

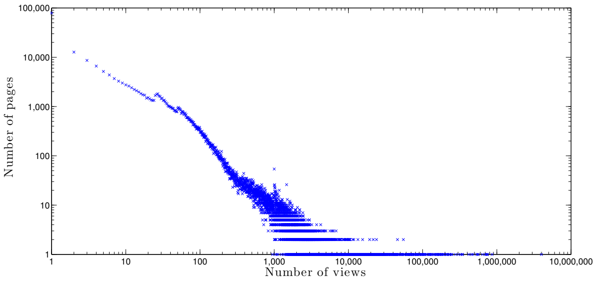

Figure 2: Log-log scatter plot of the number of views versus the number of articles in the SU index.

To give an impression of the type of articles that have gathered few or many views, Table 1 contains the ten most viewed Wikipedia articles in our data set as well as ten random examples of articles that were viewed one hundred times. We chose these two settings as they represent two extremes: on the one hand, articles that were viewed and also liked by a large number of people and on the other hand articles, that were shown a number of times but less well received by the SU users.

It should also be noted that the SU category *Bizarre & Oddities*, which dominates the list of the ten most viewed articles is not as prevalent when considering a larger set of articles. In fact, the top 100 viewed articles in our data set belong to 59 different SU categories: *Bizarre & Oddities* occurs 12 times, followed by the *Writing* category (5 times) and a number of categories with three occurrences, including *Arts*, *Science* and *Linguistics*. Only one of the top 100 articles was a so-called featured article (indicating that previous work on featured article prediction, e.g. [7], might not be applicable here), while seven were semi-protected articles due to previous vandalism activities. Notable is also the fact that 12 out of the 100 articles are of the form *List_of_X* where $X = \{$*algorithms*, *legendary_creatures*, *band_name_etymologies*$\}$ to name three examples.

While for a human reader it is usually not difficult to quickly judge whether an article is potentially interesting to him or not, it is a challenge to derive a method that automatically classifies articles accordingly. What exactly makes one article more interesting to the general public than another? In order to get get a first understanding of what users think about the most viewed articles and possibly also why they like them, we analysed the comments that were posted on the SU info page for each of the ten most viewed Wikipedia articles. This analysis is very cursory, as compared to the number of views, very few users actually comment on an article, as commenting distracts from the 'stumbling' experience. For example, the article *Wrap_rage* with 0.86 million views and forty-thousand likes has a 41 comments. In total, we analysed 479 comments and identified four broad categories:

(A) Comments expressing surprise

- "There's a name for this?"
- "I'd never heard of this before (go StumbleUpon!). Very cool."

(B) Comments expressing admiration, sadness, sorrow, etc.

- "That's so sad"
- "No one should go through life afraid to take a walk."
- "don't know what to say actually.."

(C) Comments about the usefulness of the knowledge

- "Simple, but helpful for designers."
- "An exceptional list of colours and their code, invaluable to graphic designers, webmasters etc."

(D) Comments expressing negative sentiments towards the article

- "Fake."
- "Why stumble everyday wikipedia articles?"

3.2 Working Hypotheses

Based on the preliminary qualitative insights gained, we developed three intuitions that we believe will enable us to predict to what a Wikipedia article is likely to be beneficial to the average SU user.

Intuition A. Articles that contain unexpected nuggets of information can be identified by considering how semantically related the article is to the other articles it contains links to. For instance, the *List_of_unusual_deaths* Wikipedia article has, among others, outgoing links to the following diverse articles: *Common_fig*, *Malvasia* (wine), *Eddystone_Lighthouse*, *Hawaii*, and *Chimney*. We hypothesize that finding such seemingly unrelated articles can be used as a measure of the likelihood of the article being of interest.

Intuition B. Articles that evoke emotional feelings can be discovered through a form of sentiment analysis. Although Wikipedia articles are written in a neutral style, some topics are bound to evoke emotions and those emotional topics can be identified.

Intuition C. Articles that contain useful knowledge may be identified indirectly, when considering their Talk pages, the amount of discussions that are ongoing and the style of the discussions. Articles about practically useful information are not likely to be emotionally charged, unlike discussions for instance about politicians, religious topics, etc.

We emphasize, that these are hypotheses that need to be verified in future work.

4. CONCLUSIONS

In this position paper we have proposed to investigate what makes certain Wikipedia articles interesting to users who are browsing the Web without a goal in order to pass the time or relax. Since such articles are education to some degree, the leisure activity of browsing (stumbling) can thus also incorporate a learning experience. Since life-long learning is an important characteristic of knowledge economies, it is crucial to understand the interplay between these two

Most viewed articles	#Views	#Likes	SU Category	Date	Example articles viewed 100 times	SU Category
List_of_unusual_deaths	3.99M	0.423M	Bizarre/Oddities	12/2004	Biblioscape	Software
Flying_Spaghetti_Monster	1.39M	0.121M	Satire	08/2005	Edge_of_chaos	Chaos/Complexity
Wrap_rage	0.86M	0.040M	Bizarre/Oddities	01/2008	Gottfried_Wilhelm_Leibniz_Prize	Biology
Shigeru_Miyamoto	0.75M	0.019M	Video Games	10/2003	Mario_Buda	Crime
Benjaman_Kyle	0.74M	0.051M	Bizarre/Oddities	12/2008	Proto-Indo-European_language	Linguistics
One_red_paperclip	0.72M	0.070M	Bizarre/Oddities	09/2006	Cisco_Adler	Alternative Rock
List_of_colors	0.70M	0.066M	Arts	01/2005	Biofeedback	Psychology
Do_not_stand_at_my_grave_and_weep	0.64M	0.132M	Poetry	10/2007	Ovipositor	Sexual Health
Fuel_cell	0.56M	0.009M	Science	06/2005	Concealer	Beauty
Raymond_Robinson_(Green_Man)	0.54M	0.036M	Bizarre/Oddities	05/2008	Winklepickers	Fashion

Table 1: A list of Wikipedia articles that are contained in the SU index. For the most viewed articles, shown are also the number of views and likes in million, the category in StumbleUpon the page was assigned to by the user who discovered the page and the date (month/year) at which the page was discovered.

forces. We argue that a greater understanding of features are indicative of an article's attractiveness to the average user (stumbler) will enable us to develop adaptations that expose a greater amount of Wikipedia articles to the leisure seeking user.

5. REFERENCES

[1] P. André, m. schraefel, J. Teevan, and S. T. Dumais. Discovery is never by chance: designing for (un)serendipity. In *C&C '09*, pages 305–314, 2009.

[2] P. André, J. Teevan, and S. T. Dumais. From x-rays to silly putty via uranus: serendipity and its role in web search. In *CHI '09*, pages 2033–2036, 2009.

[3] L. Björneborn. Design dimensions enabling divergent behaviour across physical, digital, and social library interfaces. In *Persuasive Technology*, volume 6137, pages 143–149. 2010.

[4] D. Bollen, B. P. Knijnenburg, M. C. Willemsen, and M. Graus. Understanding choice overload in recommender systems. In *RecSys '10*, pages 63–70, 2010.

[5] J. H. Falk and M. Storksdieck. Science learning in a leisure setting. *Journal of Research in Science Teaching*, 47(2), 2010.

[6] J. Helmes, K. O'Hara, N. Vilar, and A. Taylor. Meerkat and tuba: Design alternatives for randomness, surprise and serendipity in reminiscing. In *Human-Computer Interaction - INTERACT 2011*, volume 6947, pages 376–391. 2011.

[7] N. Lipka and B. Stein. Identifying featured articles in wikipedia: writing style matters. In *WWW '10*, 2010, pages 1147–1148.

[8] A. Oulasvirta, J. P. Hukkinen, and B. Schwartz. When more is less: the paradox of choice in search engine use. In *SIGIR '09*, pages 516–523, 2009.

[9] J. Packer. Learning for fun: The unique contribution of educational leisure experiences. *Curator: The Museum Journal*, 49(3):329–344, 2006.

[10] J. Packer. Beyond learning: Exploring visitors' perceptions of the value and benefits of museum experiences. *Curator: The Museum Journal*, 51(1):33–54, 2008.

[11] J. Packer and R. Ballantyne. Motivational factors and the visitor experience: A comparison of three sites. *Curator: The Museum Journal*, 45(3):183–198, 2002.

[12] J. M. Packer. *Motivational factors and the experience of learning in educational leisure settings.* PhD thesis, Queensland University of Technology, 2004.

[13] G. Wu, M. Harrigan, and P. Cunningham. Characterizing wikipedia pages using edit network motif profiles. In *SMUC '11*, pages 45–52, 2011.