Mining for Social Serendipity

Alexandre Passant[1], Ian Mulvany[2], Peter Mika[3], Nicolas Maisonneauve[4], Alexander Lser[5], Ciro Cattuto[6], Christian Bizer[7], Christian Bauckhage[8], Harith Alani[9]

[1] DERI, National University of Ireland, Galway - alexandre.passant@deri.org
[2] Nature Publishing Group - ian@mulvany.net
[3] Yahoo Research, Barcelona - pmika@yahoo-inc.com
[4] Sony CSL, Paris - n.maisonneuve@gmail.com
[5] TU Berlin - aloeser@cs.tu-berlin.de
[6] ISI Foundation - ciro.cattuto@roma1.infn.it
[7] FU Berlin - christian.bizer@fu-berlin.de
[8] Deutsche Telekom Laboratories - christian.bauckhage@telekom.de
[9] University of Southampton - ha@ecs.soton.ac.uk

Abstract. A common social problem at an event in which people do not personally know all of the other participants is the natural tendency for cliques to form and for discussions to mainly happen between people who already know each other. This limits the possibility for people to make interesting new acquaintances and acts as a retarding force in the creation of new links in the social web. Encouraging users to socialize with people they don't know by revealing to them hidden surprising links could help to improve the diversity of interactions at an event. The goal of this paper is to propose a method for detecting *"surprising"* relationships between people attending an event. By *"surprising"* relationship we mean those relationships that are not known a-priori, and that imply shared information not directly related with the local context of the event (location, interests, contacts) at which the meeting takes place. To demonstrate and test our concept we used the Flickr community. We focused on a community of users associated with a social event (a computer science conference) and represented in Flickr by means of a photo pool devoted to the event. We use Flickr metadata (tags) to mine for user similarity not related to the context of the event, as represented in the corresponding Flickr group. For example, we look for two group members who have been in the same highly specific place (identified by means of geo-tagged photos), but are not friends of each other and share no other common interests or, social neighborhood.

Key words: serendipity, online activity, context, ubiquitous computing

1 Motivation

In the course of discussion at the workshop a theme which this group came to was to question the purpose of the applications that we make. We were led to ask whether we could devise an application, based on available information

Dagstuhl Seminar Proceedings 08391
Social Web Communities
http://drops.dagstuhl.de/opus/volltexte/2008/1791

from the social web, that could in some way increase the delight that a person experienced in their life. In thinking about what causes delight we quickly focused on the idea of surprise, and in particular serendipitous surprises. Wikipedia defines serendipity as *"the effect by which one accidentally discovers something fortunate, especially while looking for something else entirely"*[10]. Julius Comroe described it as *"looking in a haystack for a needle and discovering a farmers daughter"*. A well known example is the discovery of Penicillin by Alexander Fleming. He failed to disinfect cultures of bacteria when leaving for vacation, only to find them contaminated with Penicillium molds, which killed the bacteria. He had previously done extensive research into antibacterial substances, and so this discovery had a particular resonance for him. To be sure, surprise is not always a good thing. An example is the invention of Gelignite by Alfred Nobel, when he accidentally mixed collodium (gun cotton) with nitroglycerin. Though initially a good accident the personal regret that he experienced through the misuse of his invention led to the creation of the foundation that now bears his name.

In a social situation, for instance a topic-oriented meeting, intuitively it is not surprising that there are common connections between people, and that people who share those connections socialize together. Those connections can be purely social graph connections (they know people in common), geo-historical connections (they have been to similar events in the past) or connections of interest (they have published in the same sub-field). Even if two attendees at a conference are unaware of their shared trips to similar conferences in the past, such a connection is not really surprising as it may be inferred from the shared context of the current meeting. Intuitively these connections are not really surprising. It is more surprising if two people shared a hidden relationship which is clearly not related to the topic of the shared context. Such a connection may be even even more surprising if that relationship is not common to other members of the shared group. If such relationships, or connections, in the extended social graph, or the social hyper-graph, could be uncovered, or mined prior to, or even at an event, they may form a sufficient catalyst for people at the event to forge new relationships or connections.

2 Approach

2.1 The Social Hyper-graph

As Web 2.0 sites and social networking sites expand the number of traces that people leave in the web grows. Traces of social connections have been available for a long time, the co-authorship network of academic literature is one much studies example, however as threshold for what constitutes a signal decreases the volume of information available increases correspondingly. The poster boy of the Web 2.0 world in 2008 has been the microblogging service Twitter [11], that lead to ubiquitous life streaming, especially as people can update their account

[10] `http://en.wikipedia.org/wiki/Serendipity` - Accessed on 22nd Oct. 2008
[11] `http://www.twitter.com` - Accessed on 22nd Oct. 2008

using their mobile phone, but no less important is the emerging web of things. Application layers such as Yahoo! Research's Fire Eagle make possible the creation of social networks that are aware of the location of their members in real time[12]. Sites such as Doppler [13] already use this kind of information to mediate connections between people [14]. [15]. Of emerging interest will be the traces left by devices embedded in the environment. Mobile phones in particular already provide a platform for embedded sensors. The traces left by these devices are leaving nascent connections between not just people, but between people, places and things. The type of graph that best describes these relationships will probably be some kind of an n-partate or hyper graph, connecting on-line and off-line worlds. There is considerable scope for investigating which of the relationships that may emerge from this kind of data constitute important relationships, but in this paper we try to outline an approach to uncovering surprising relationships in any given network.

2.2 Surprise is an Attribute.

A measure of surprise can only be considered in relation to the item which is described as surprising, e.g. an event, a TV show, a person. Surprise can modeled through an event based approached using Baseian statistics. By comparing the difference between the prior and posterior probablitliy of an event happening a measure for the surprisingnes of the event can be generated. Following directly the description of surprise given in the Formal Bayesian Theory of Surprise[16] we can set a measure for the surprise S of a the event or observation of interest D given the model space $\mathcal{M}$ of the observer. This is given as the relative entropy of the Kulback-Leibler divergence of prior and posterior probability distribution of the likelihood of the event happening. What we are doing is measuring how far apart the prior and posterior distributions are. When they are far apart the event is surprising and it's measure is given by:

$$S(D, \mathcal{M}) = \mathrm{KL}(P(M|D), P(M)) = \int_{\mathcal{M}} P(M|D) \log \frac{P(M|D)}{P(M)} \mathrm{d}M.$$

In the context of social web communities the event in question will be the creation of a link in some network. The key question in regard to this work is figuring out how to model the $\mathcal{M}$ that an observer, or user, will have of the

[12] `http://feblog.yahoo.net/2008/07/17/being-social-with-fire-eagle/` - Accessed on 11th Nov. 2008

[13] `http://www.slideshare.net/blackbeltjones/reboot90-travel-serendipity` - Accessed on 11th Nov. 2008

[14] `http://www.slideshare.net/blackbeltjones/reboot90-travel-serendipity` - Accessed on 11th Nov. 2008

[15] `http://www.zephoria.org/thoughts/archives/2008/09/21/i_will_be_joini.html` - Accessed on 11th Nov. 2008

[16] `http://ilab.usc.edu/surprise/` - Accessed on 14th Nov. 2008

explicit links that they have in their network. By uncovering a hidden link that they may not be aware of, we change the structure of the network of explicit links of the observer. Which of the many threads that connect us to each other will bring about the greatest change in the way that the observer perceives the links they already know about.

2.3 Serendipity is a Surprising Event.

As stated, our goal in this paper is to describe and propose a system which increases the serendipidous happenings in a users life. We can expect that a typical event at a conference is that people will meet each other. Can we provide a system which uncovers the meetings that would be maximally surprising for the people involved? The prior expectation for people at a meeting is that they will have nothing in common outside of the shared topic of the meeting with other members of the meeting. By suggesting connections with people who have a strong, but unknown connection between each other we may broker surprising meetings.

More formally, we define a *surprising relationship* as a relationship between two individuals (e.g. Alice and Bob) having the following properties:

- Alice and Bob do not know each other from before, hence do not share a common edge in their relationship network, whether it is online or in the real-world;
- The unexpected shared interest is not related to the context (time, location, friends ...) they share, and neither is it common to other members of the community;

Basically, our system finds unexpected answsers to the question *"What topic do I share with X?"*. The more the topic is unrelated to the context in which the question is asked, the bigger the surprise is.

2.4 A Simple Model

We propose to create a representation of the model of a user, or a user's interests, based on the tags that they leave in social web applications. Hence, we bridge off-line and on-line world, by identifying relationships between people based on their past footprints on the Web 1. Ideally one would wish to make a more refined model through utilizing as much information as was available, and appropriate for use, however for a first pass it is instructive to try with the an easily available data set. By taking a simple distance metric of the vectors created by a users personomy we can find users from a group who may be close or far from each other in the context of a given interest. By looking for personomy information that may be outside of the information common to a group, the tags of an event from Flickr, for instance, we can look for connections that might be maximally surprising thanks to what people tagged in the past. Key to this is that the metric over which we search is not a simple additive space in which we try

to minimize the distance between all of the available sources of information, but rather that we also try to find at least some attributes over which we find a maximal distance, so that the eventual measure we come up with has the potential to reveal surprising connections.

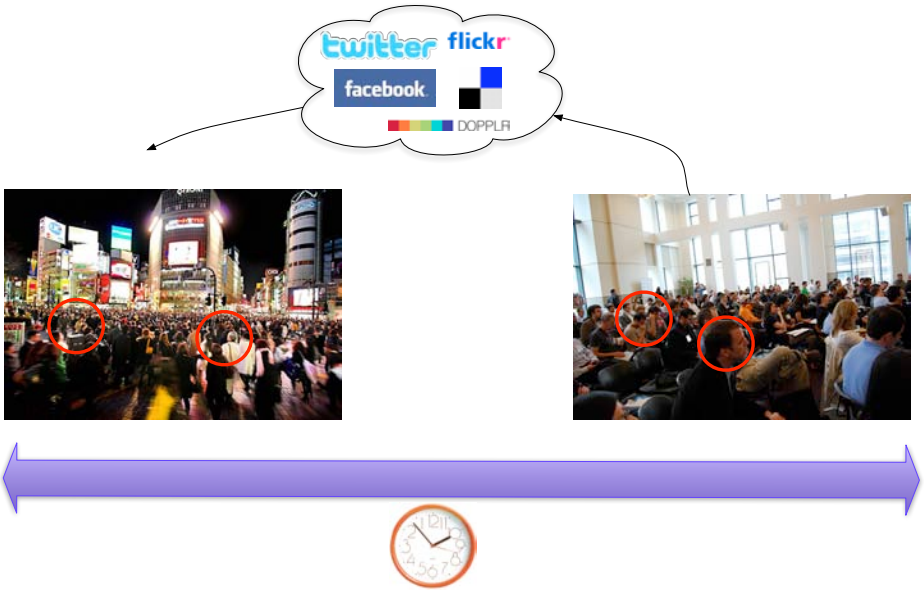

Fig. 1. Mining surprise thanks to online footprints from the past

3 Related Work

In this section we briefly introduce to existing research from the area of recommender systems, web mining and previous work on theories of surprise.

3.1 Measuring Serendipity in Recommender Systems

In [2], the authors propose metrics *unexpectedness* and *unexpectedness_r* for measuring the serendipity of recommendation lists produced by recommender systems. In their work *unexpectedness* is a metric for a whole recommendation list, while *unexpectednesss_r* is the metric that takes into account the ranking in the list.

3.2 Web Mining

The authors of [1] proposes a new method for answering relationship queries on two entities e.g., the connections between different places or the commonalities of people. Their method matches top Web pages retrieved for individual entities and automatically identifies the connecting terms. To effectively filter out the large amount of noise in the Web pages without losing much useful information, we do windowing around query keywords, compute term weights based on the characteristics of the two Web page sets, and only use the top potential connecting terms to compute the similarity values of Web page pairs. Their work is orthogonal, since they are trying to minimize the distance between the concepts requested in the relationship query. However, we are trying to maximize the distance in a subset of the data avaliable about the query with the aim of maximizing the surprise the result. The authors of [3] extract the location and event data from Flickr data by usage distribution of each tag. They analyze mappings of location and time metadata associated with photos and their tags and apply different distance metrics.

4 Metrics for Evaluating the Serendipity of Recommendation Lists

4.1 Experiment

In order to evauluate or approach, we designed the following algorithm.

- Extract the tag distribution of the event based on the photos in the group (E) from Flickr;
- For each user:
  - Compute the difference of the event distribution (E) and the individual distributions ($U1, U2..$);
  - Store this as non event related interest ($UNE1, UNE2$);
- Extract the tag distribution of all members in the group (G);
- For each user:
  - Compute the difference of (G) and the individual distributions;
  - Store this as non group related interest ($UNG1, UNG2$);
- For each pair of users:
  - Compute the correlation of their non event related interests or at least find one strong shared interest that are not common in U
- Another example is to have an augmented experience of a place by showing to the user his 'connection', e.g a friend of him has visited this place recently. to diversify the interactions

4.2 Results

We ran that algorithm on a dataset of more than 900 pictures and around 40 people from a Flickr group related to the ISWC conference series [17]. We indeed identified two related users by shared topics that were not related to the event topics. Unfortunately, it turned out that those two people were colleagues from the same institution, however our algorithm was quite basic and did not take into account even simply available information, such as the two individuals in question being connected in Flickr by friendship.

5 Extending the Surprise

Based on our definition of social serendipity and surprising relationship, we identified several factors that might extend the strength of the surprise. Indeed, some relationships are stronger that others. For instance, identifying that two people in an event are the only two to have visited the Effeil Tower might be relevant from the conference context point of view, but not at all from a global overview, as the Effeil Tower is quite a famous location. Far more surprising would be had those two people been the only two at the conference to have visited the Eiffel Tower french restaurant in Burnley.

5.1 Tag Specificity

Obviously, some tags are most restrictive that others. While our current approach takes only the context into account to identify related tags, we must also consider a global approach to identify really interesting tags. To achieve this goal, we can rely on different strategies:

- Considering the current social networking website that we analysed, a less time a tag has been used in the whole service, the most relevant it is. Regarding Flickr, we can add a strongest weight to the tags that have been used only a few times;
- We can also extend the approach to use background knowledge. Ranking by inverse number of results in search engines might also re-enforce the value of the tag specifity;
- Wikipedia might also be used to identify specific tags. Here, we can consider the fact that a page exists for that tag or not, but also its volume and its activity. A page edited by only a few people certainly identify a niche topic while a page edited by 100s of people is obviously related to a well-known topic. In case a Wikipedia page is found, but not in any language, it might also be a factor to identify a niche topic, i.e. foreign people that are interested in a partucular aspect of a country's culture which is not known abroad:

[17] `http://flickr.com/groups/iswc/` Accessed on 14th Nov. 2008

5.2 Semantic Web and Tags Relationships

Instead of considering tag identity, we may also deal with related tag, thanks to similarity measures and semantic distance. Bridging the gap between the tags and the meaning they represent, modeled thanks to Semantic Web principles and knowledge bases like DBpedia[18], or more generally, data from the Linking Open Data project[19], would provide a large base of interlinking concepts. Then, we should compute some semantic distance measures to identify related tags, e.g. a tag identifying a band and another one related to the name of its singer.

5.3 Interlinking Social Networks and Folksonomis

Our current approach relies on a single Social Media platform to find unexpected relationships between people. To go further, we should consider the different social websites a user is member of, and consequently its whole social activity. With the growth of Web2.0, it is becoming increasingly common for users to maintain a presence in more than one site. For example, one could be bookmarking pages in del.icio.us, uploading images in Flickr, listening to music in Last.fm, blogging in Technorati, etc. The nature of these pursuits naturally leads users to express the relevant aspects of their interests, which are likely to be different across the sites. If such multiple identities and distributed activities could be brought together independently of the Web 2.0 sites thanks to the previously introduced models, far better understanding can be reached about what users are interested in.

Many popular sites are racing to develop tools to allow their users to carry their personal profiles to other sites. Within days from each other in May 2008, Google, MySpace, and Facebook announced new initiatives for increasing social profile portability called Friend Connect[20], Data Availability, and Connect[21] respectively. Yet, those projects remain efforts of private companies and might not be able to be linked each other for economic reasons. On the other end, efforts from the Semantic Web community such as FOAF[22] or SIOC[23] might be considered to ease that process. They provide a common semantics and related tools to interlink distributed social networks and related data that will lead to distributed, open and standard-based social graph modeling. Various applications providing those data from existing services have been build, for instance exporters for Flickr[24] of Facebook[25], as well as exporters for major blogging platforms[26].

[18] `http://dpedia.org` Accessed on 14th Nov. 2008
[19] `http://linkeddata.org` Accessed on 14th Nov. 2008
[20] `http://www.google.com/friendconnect/` Accessed on 14th Nov. 2008
[21] `http://developers.facebook.com/connect.php` Accessed on 14th Nov. 2008
[22] `http://foaf-project.org` Accessed on 14th Nov. 2008
[23] `http://sioc-project.org` Accessed on 14th Nov. 2008
[24] `http://apassant.net/home/2007/12/flickrdf/` - Accessed on 14th Nov. 2008
[25] `http://www.dcs.shef.ac.uk/~mrowe/foafgenerator.html` - Accessed on 14th Nov. 2008
[26] `http://sioc-project.org/applications` - Accessed on 14th Nov. 2008

Cross-folksonomy integration will become the focus of much research and development in the near future. There is a strong push towards opening up social networking to support portability of data across various sites. . Folksonomies can be interlinked at multiple levels:

- User Correlation: Since the same individual may hold accounts in multiple social networking sites, two separate folksonomies may be joined at the user level. Such a joining enables one to analyse and study the tagging practices of individuals across different tagging platforms.
- Tag Correlation: It is likely that many tags will appear in more than one folksonomy. As a result, cross-folksonomy networks can be generated by joining the common tags.
- Resource Consolidation: Finally, the resources themselves may appear across multiple folksonomies. By understanding how resources in different folksonomies relate to each other, e.g. through shared tags or users, it would be possible to provide a consolidated view of resources distributed over multiple folksonomies.

The SCOT project[27] aims to achieve this goal of tag portability by providing a model - and related tools - for tagging actions using Semantic Web technologies.

5.4 From tags to interest

Tags serve various purposes, such as for resource organisation, promotions, sharing with friends, with the public, etc. However, studies have shown that tags are generally chosen to reflect their user's interests. If we can distill users' Web 2.0 activities to produce an accurate image of their interests, then far better identification of surprising links could be achieved.

Tag combinations Moving from a cloud of tags to a list of interests is not trivial. Not all tags reflect an interest. For example, when we examined some del.icio.us users, we found one person who tagged a resource with "time", which is an overgeneral tag, and hence not very helpful for learning about the user's interests. However, when we looked at other tags for this user on comparable posts, it became clear that the particular interest of the user was astronomy, in which the concept of time plays a specific role.

Tag disambiguation Dynamic tag disambiguation is another challenging problem. Some methods have been suggested, which are based on clustering the whole collection of tags or resources, but such techniques are more suitable for static environments where the data do not change too often. In the world of tagging, this is hardly the case. Hence we need less demanding methods to disambiguate tags, to cope with the highly dynamic nature of folksonomies, even if those methods could never be perfect

[27] `http://scot-project.org` Accessed on 14th Nov. 2008

Tag cleaning Tags are free text, and users can tag resources with any terms they wish to use. On the one hand, this total freedom simplifies the process and thus attracts users to contribute. It also avoids the problem of forcing users into using terms they do not feel apply, as opposed to enforcing the use of a set of terms. For these reasons, the lack of constraints seems essential. On the other hand, it generates various vocabulary problems, where tags can be too personalised, made of compound words, mix plural and singular terms, meaningless, synonymous, etc. To make tags more comparable, it will be necessary to deploy a number of tag cleaning processes to increase their compatibility, which in return should increase the quality of interest identification.

5.5 Temporal Factor

Finally, we can extend our model by taking the temporal aspect into account. While a tagging action is usually represented as a tripartite graph, most websites store the timestamp of that association. By considering this additional dimension, we should be able to identify tags that user shared at a given time which can help to identify people sharing a common interest at the same time, or, considering geolocation tag, people that have been in the same place at the same time. This last example shows that our method can help to bridge the gap between real world and online activities. More specifically, we identify real-world relationships (based on geolocation), thanks to online activities, thus closing a triangle combining real-world and online representation. Of course, this temporal factor could be combined with the previous semantic distance, identifying facts that two people were at the same time in two small pubs in a suburb of Tokyo two years ago. Considering simple non-geographic tags, that temporal aspect can also be used to identify early-adopters or experts in a given field, using services as google trends of web archives.

5.6 Geographical Factor

By looking for geotags we can get an idea of the spatial area that the person has visited. By correlating an area by it's size and by it's 'fame', as perhaps derived from Wikipedia, we can weight those geotags accordingly. Smaller less famous locations being more likely to lead to a surprising relationship.

5.7 Applications

We can envisage many application but two immediate applications that present themselves are in the realms of dating services and scientific collaboration. By being able to provide an ice-breaking connection between two people we imagine that such a system would enable people to make connections faster. By looking at the co-authorship network such a system may be able to recommend interesting cross discipline partnerships. By finding people who are close in one metric (a technique say), but distant in another metric (a field say) such a system could aid researchers by providing paths out of the local minima that they work in.

6 Further Work

The implementation that we developed was clearly a toy implementation. It may be of interest to take forward some of the ideas surrounding the surprisingness of signals (astonishment value). A key issue is trying to develop more formally the relationship between a personomy and the model that a user might have about the world around them. When we begin to measure distances over a space in which we try to maximize our result over some dimensions and minimize our result over other dimensions questions about the concavity of the space may arise, in addition to questions about finding an optimal search strategy.

7 Conclusion and Comments

By thinking in a somewhat unexpected direction about the information that we can infer from the electronic traces and detritus that people leave in their electronic wakes, we have found a novel and potentially interesting methodology for uncovering surprising relationships, however it should be noted that not all things that are surprising are by default a good thing. Discovering that someone whom one had expected to be similar actually holds radically different views from oneself can lead initially to surprise and then could lead to discomfort. We term the difference between *good* surprise and *bad* surprise the difference between the uncanny and the unexpected. A significant challenge for any system that implements maximal surprise is to distinguish between good surprise and bad surprise.

References

1. Gang Luo, Chunqiang Tang, and Ying li Tian. Answering relationship queries on the web. In Carey L. Williamson, Mary Ellen Zurko, Peter F. Patel-Schneider, and Prashant J. Shenoy, editors, *WWW*, pages 561–570. ACM, 2007.
2. Tomoko Murakami, Koichiro Mori, and Ryohei Orihara. Metrics for evaluating the serendipity of recommendation lists. In Ken Satoh, Akihiro Inokuchi, Katashi Nagao, and Takahiro Kawamura, editors, *JSAI*, volume 4914 of *Lecture Notes in Computer Science*, pages 40–46. Springer, 2007.
3. Tye Rattenbury, Nathaniel Good, and Mor Naaman. Towards automatic extraction of event and place semantics from flickr tags. In *SIRIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 103–110, New York, NY, USA, 2007. ACM Press.