Introducing Serendipity in a Content-based Recommender System

Leo Iaquinta, Marco de Gemmis, Pasquale Lops, Giovanni Semeraro,
Michele Filannino, Piero Molino
Dipartimento di Informatica, Università degli studi di Bari – Italia
{iaquinta, degemmis, lops, semeraro}@di.uniba.it
info@filanninomichele.com, piero.molino@gmail.com

Abstract

Today recommenders are commonly used with various purposes, especially dealing with e-commerce and information filtering tools. Content-based recommenders rely on the concept of similarity between the bought/searched/visited item and all the items stored in a repository. It is a common belief that the user is interested in what is similar to what she has already bought/searched/visited. We believe that there are some contexts in which this assumption is wrong: it is the case of acquiring unsearched but still useful items or pieces of information. This is called serendipity. Our purpose is to stimulate users and facilitate these serendipitous encounters to happen.

This paper presents the design and implementation of a hybrid recommender system that joins a content-based approach and serendipitous heuristics in order to mitigate the over-specialization problem with surprising suggestions.

1. Background and Motivation

Information overload is a common issue among the modern information society. Information Filtering (IF) is a kind of intelligent computing techniques that mitigates this problem by providing the user with the most relevant information with respect to her information needs.

Recommender systems (RSs) adopt IF techniques in order to provide customized information access for targeted domains.

They can be viewed as intelligent systems that take input directly or indirectly from users and, based on their needs, preferences and usage patterns, provide personalized advices about products or services and can help people to filter useful information.

Several definitions of RS have been given. According to [3]: "Recommender systems have the effect of guiding the user in a personalized way to interesting or useful objects in a large space of possible options". This definition makes it clear that user oriented guidance is critical in a RS.

Among different recommendation techniques proposed in the literature, the content-based and the collaborative filtering approaches are the most widely adopted to date. Systems implementing the content-based recommendation approach analyze a set of documents, usually textual descriptions of the items previously rated by an individual user, and build a model or profile of user interests based on the features of the objects rated by that user [14]. The profile is exploited to recommend new items of interest. Collaborative recommenders differ from content-based ones in that user opinions are used instead of content. They gather ratings about objects by users and store them in a centralized or distributed database. To provide user X with recommendations, the system computes the neighborhood of that user, i.e. the subset of users that have a taste similar to X. Similarity in taste is computed based on the similarity of ratings for objects that were rated by both users. The system then recommends objects that users in X's neighborhood indicated to like, provided that they have not yet been rated by X. Each type of filtering methods has its own weaknesses and strengths.

In particular, the content-based approach suffers from *over-specialization*. When the system can only recommend items that score highly against a user's profile, the user is limited to being recommended items similar to those already rated. Even a 'perfect' content-based technique would never find anything surprising, limiting the range of applications for which it would be useful. This shortcoming is called *serendipity problem*.

To give an example, a person with no experience with Greek cuisine would never receive a recommendation for even the greatest Greek restaurant in town.

In other words, over-specialized systems can prevent serendipitous discoveries to happen, according to Gup's theory [9].

It is useful to make a clear distinction between *novelty* and *serendipity*. As explained by Herlocker

978-0-7695-3326-1/08 $25.00 © 2008 IEEE
DOI 10.1109/HIS.2008.25

[10], novelty occurs when the system suggests to the user an unknown item that she might have autonomously discovered. A serendipitous recommendation helps the user to find a surprisingly interesting item that she might not have otherwise discovered (or it would have been really hard to discover). To provide a clear example of the difference between novelty and serendipity, consider a recommendation system that simply recommends movies that were directed by the user's favorite director. If the system recommends a movie that the user was not aware of, the movie will be novel, but probably not serendipitous. On the other hand, a recommender that recommends a movie by a new director is more likely to provide serendipitous recommendations. Recommendations that are serendipitous are by definition also novel.

Novelty is the main objective of a "classical" recommender. We agree with the theory proposed by McNee [12], that studies about improving precision and recall (or accuracy metrics in general) just do not get the point of what is useful for the user: a sensible recommendation (which is not always the most accurate one).

Our objective is to try to feed the user also with recommendations that could possibly be serendipitous. In this paper, we suggest a possible way to introduce serendipity within a recommender system or in a generic digital library. We will show how the architecture of content-based RS might be extended in order to introduce a component devoted to introduce serendipity in the recommendation process in an operational way.

The paper is organized as follows: Section 2 presents the serendipity issue for information seeking; Section 3 covers strategies to provide serendipitous recommendations; Section 4 provides a description of our recommender system and how it discovers potentially serendipitous items in addition to content-based suggested ones; Section 5 provides the description of the experimental session carried out to evaluate the proposed ideas; finally, Section 6 draws conclusions and provides directions for future work.

2. Serendipity and information seeking

Horace Walpole coined the term "serendipity" in the 1754 explaining it as "*making discoveries by accident and sagacity of things which one is not on quest of*". The origin of the word "serendipity" [18] is the persian fairy tale titled "The three princes of Serendip" that Cristoforo Armeno translated and published in the 1557. M.K. Stoskopf [8] was one of the first scientists to acknowledge the relevance that serendipity covers in scientific field, affirming that serendipitous discoveries are of significant value in the advancement of science and often presents the found for important intellectual leaps of understanding.

The history of science is full of serendipitous discoveries: the (re-)discovery of the Americas by Columbus, the Gelignite by Nobel, the Penicillin by Fleming, etc.

We agree with Roberts [15] when he stresses that serendipitous encounters depend on the characteristic of the information seeker, her open minded attitude, her wide culture and her curiosity.

The idea of serendipity has a link with de Bono's "lateral thinking" [6] which consists not to think in a selective and sequential way, but accepting accidental aspects, that seem not to have relevance or simply are not sought for. This kind of behavior surely helps the awareness of serendipitous events.

The subjective nature of serendipity is certainly quite a problem when trying to conceptualize, analyze and implement it. As Foster & Ford said: "*Serendipity is a difficult concept to research since it is by definition not particularly susceptible to systematic control and prediction. […] Despite the difficulties surrounding what is still a relatively fuzzy sensiting concept, serendipity would appear to be an important component of the complex phenomenon that is information seeking*" [8]. Even though we agree with van Andel [18] that we cannot program serendipity because of its nature, we share the concern of Campos and de Figueiredo [4] of programming for serendipity. They also tried to suggest a formal definition of serendipity [5] identifying different categories for serendipitous encounters.

By the way, the problem of programming for serendipity has not been deeply studied and there are really few theoretical and experimental studies.

The noble objective of allowing users expand their knowledge and preserve the opportunity of making serendipitous discoveries even in the digital libraries could push the development of useful tools that can facilitate important intellectual leaps of understanding.

Like Toms explains [17], there are three kind of information searching:

- seeking information about a well-defined object;
- seeking information about an object that cannot be fully described, but that will be recognized at first sight;
- acquiring information in an accidental, incidental, or serendipitous manner.

It is easy to realize that serendipitous happenings are quite useless for the first two ways of acquisition, but are extremely important for the third kind.

As our work concerns the implementation of a serendipity–inducing module for a content-based

recommender, the appropriate metaphor in a real-world situation could be one of a person going for shopping or visiting a museum who, while walking around seeking nothing in particular, would find something completely new that she has never expected to find, that is definitely interesting for her.

3. Strategies to induce serendipity

We have the problem of introducing serendipity in the recommendation process in an operational way. Among different approaches which have been proposed, Toms suggests four strategies, from simplistic to more complex ones [17]:

1. Role of chance or 'blind luck', implemented via a random information node generator.
2. Pasteur principle ("chance favors the prepared mind"), implemented via a user profile.
3. Anomalies and exceptions, partially implemented via poor similarity measures.
4. Reasoning by analogy, whose implementation is currently unknown.

In this work we propose an architecture for content-based RSs that implements the "Anomalies and exceptions" approach, in order to provide serendipitous recommendations alongside classical ones, thus providing the user with new entry points to the items in the system.

The basic assumption is that serendipity cannot happen if the user already knows what is recommended to her, because a serendipitous happening is by definition something new. Thus the lower is the probability that user knows an item, the higher is the probability that a specific item could result in a serendipitous recommendation. The probability that user knows something semantically near to what the system is confident she knows is higher than the probability of something semantically far. If we evaluate semantic distance with a similarity metric, like internal product which takes into account the item description to build a vector and compares it to other item vectors, it results that it is more probable to get a serendipitous recommendation providing the user with something less similar to her profile.

According to this idea, items should not be recommended if they are too similar to something the user has already seen, such as different news article describing the same event.

Therefore, some content-based RSs, such as DailyLearner [2], filter out items not only if they are too different from the user preferences, but also if they are too similar to something the user has seen before. Following this principle, the basic idea underlying the proposed architecture is to ground the search for potentially "serendipitous" items on the similarity between the item descriptions and the user profile, as described in the next section.

4. Inducing serendipity in a content-based recommender

ITem Recommender (ITR) is a content-based recommender system, developed at the University of Bari [7] [16].

The system is capable of providing recommendations for items in several domains (e.g., movies, music, books), provided that descriptions of items are available as text documents (e.g. plot summaries, reviews, short abstracts).

In the following, we will refer to documents as textual descriptions of items to be recommended.

Figure 1 shows the general architecture of the system. The recommendation process is performed in three steps, each of which is handled by a separate component.

4.1. Content Analyzer

It allows introducing semantics in the recommendation process by analyzing documents in order to identify *relevant concepts* representing the content. This process selects, among all the possible meanings (senses) of each polysemous word, the correct one according to the context in which the word occurs. In this way, documents are represented using concepts instead of keywords, in an attempt to overcome the problems due to natural language ambiguity. The final outcome of the preprocessing step is a repository of disambiguated documents. This semantic indexing is strongly based on natural language processing techniques and heavily relies on linguistic knowledge stored in the WordNet lexical ontology [13].

The core of the Content Analyzer is a procedure for Word Sense Disambiguation (WSD), called JIGSAW [1]. WSD is the task of determining which of the senses of an ambiguous word is invoked in a particular use of that word. The set of all possible senses for a word is called *sense inventory* that, in our system, is obtained from WordNet. The basic building block for WordNet is the *synset* (*syn*onym *set*), which contains a group of synonymous words that represents a concept. Since it is not the focus of the paper, the procedure is not described here. What we would like to underline here is that each document is indexed as a list of WordNet synsets, thus shifting the document representation from keywords to concepts in the WordNet ontology, i.e. the synsets.

4.2. Profile Learner

It implements a supervised learning technique for learning a probabilistic model of the interests of the active user from disambiguated documents rated according to her interests. This model represents the semantic profile, which includes those concepts that turn out to be most indicative of the user preferences.

We consider the problem of learning user profiles as a binary Text Categorization task, since each document has to be classified as interesting or not with respect to the user preferences. Therefore, the set of categories is restricted to POS, that represents the positive class (user-likes), and NEG the negative one (user-dislikes). The induced probabilistic model is used to estimate the a-posteriori probability, $P(X|d)$, of document d belonging to class X.

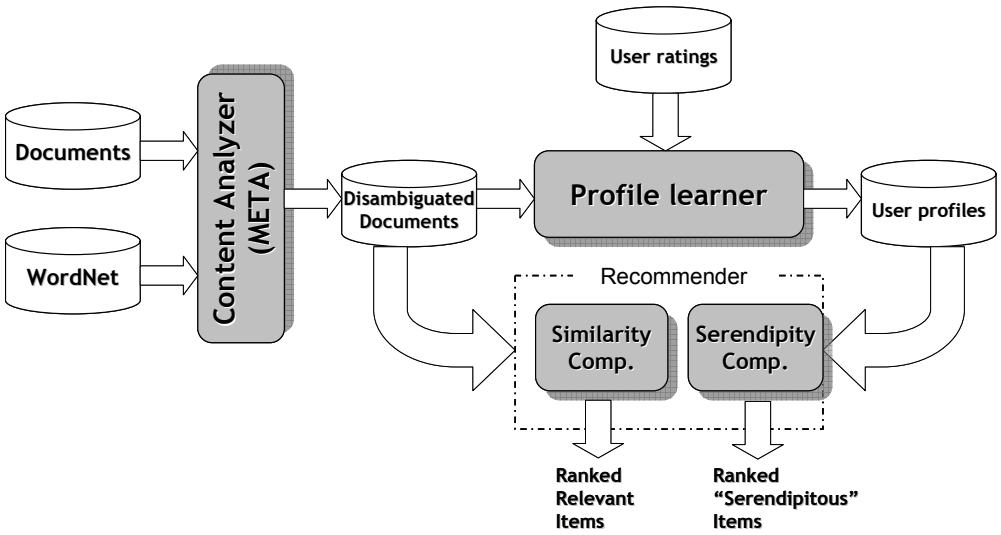
The algorithm adopted for inferring user profiles is a Naive Bayes text learning approach, which is not fully described here due to space limitations. More details are reported in [16]. What we would like to point out here is that the final outcome of the learning process is a text classifier able to categorizes a specified item in two classes: POS (for the item the user should like) and NEG (for the item the user should not like). The classifier is inferred by exploiting items labeled with ratings from 0 to 5 (items rated from 0 to 2 are used as training examples for the class NEG, while items rated from 3 to 5 are used as training examples for POS).

4.3. Recommender

It exploits the user profile to suggest relevant documents, by matching concepts contained in the semantic profile against those contained in documents to be recommended. The module devoted to discover potentially serendipitous items has been included in this component, in addition to the module which is responsible for the similarity computation between items and profiles.

In order to integrate Toms' "poor similarity" within the recommender, a set of heuristics has been included in the module for *serendipity computation*. The module devoted to compare items with profiles (Similarity Computation) produces a list of items ranked according to the a-posteriori probability for the class POS. That list will contain on the top the most similar items to the user profile, i.e. the items high classification score for the class POS. On the other hand, the items for which the a-posteriori probability for the class NEG is higher, will ranked lower in the list. The items on which the system is more uncertain are the ones for which difference between the two classification scores for POS and NEG tends to zero. We could reasonably assume that those items are not known by the user, since the system was not able to clearly classify them as relevant or not. Therefore, one of the heuristics included in the serendipity module takes into account the absolute value of the difference of the probability of an item to belong to the two classes: $|p(\text{POS}|d)-p(\text{NEG}|d)|$. The items for which the lowest difference $|p(\text{POS}|d)-p(\text{NEG}|d)|$ is observed is the

Figure 1. General system architecture



most uncertainly categorized, thus it might result to be the most serendipitous one.

5. Experimental Session

The experimentations we conducted was based over a corpus of 45 paintings chosen from the collection of the Vatican picture-gallery. Each item in the dataset has an image and three textual metadata (title, artist, and description). Figure 2 shows a sample.

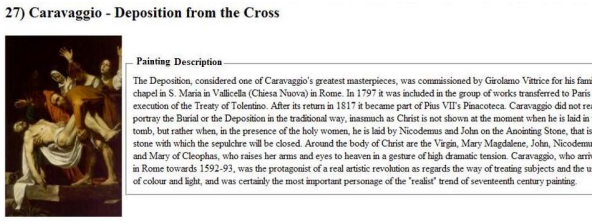


Figure 2. Sample of dataset item

We involved 30 users who voluntarily took part in the experiments. The average age of the users was in the middle of twenties. None of the users was an art critic or expert. Users were requested to express their preferences for collection items as a numerical vote on a 5-point scale (1=strongly dislike, 5=strongly like).

The ratings have been used for K-fold cross validation [11] that gave back an average degree of precision, recall and F-measure [10]. We simulated the user interaction with the system using a small part of the ratings of each user for the training of the classifier. Then five iterations were performed, in which a serendipitous item was selected by the module, rated with the ratings already expressed and added to the training set. The ratings for 5 serendipitous items proposed to each one of the 30 users were collected. The whole process has been done with the most serendipitous function and for the random over most serendipitous function with a numeric threshold of 5%, 10% and 15% of the database.

The simulating approach results from the goal to investigate different strategy for providing serendipitous recommendation. The rating interpretation issue comes out. Indeed, from a pragmatic point of view, ratings must be homogenous with other ratings in order to allow their subsequent exploitation in the profile learning step, but user rating motivations affect the meaning evaluation of finding unknown and possibly interesting things, and not simply interesting ones. For instance, a poorly rating for serendipitous suggested items should come from the experience of the user (the user already knows the item), from her lack of interest (the user already knows the item and is not interested in it), from her lack of interest in finding new things (the user does not know the item and has no interest in knowing something new), from the conscious expression of dislike (the user did not know the item before, now she knows it but she does not like it or is not interested in it) or from a serendipitous encounter (before-unknown item that results to be interesting for the user).

The results of the experimentation showed that the average percentage of c^+ items (ratings better than 3) were 40,67% for the most serendipitous function, 42,67% for the random over most serendipitous function with a threshold of 5% of the database size, 46,67% with a 10% threshold and 48,67% with a 15% threshold.

Table 1. Four functions average results

	Average c^+
Most serendipitous	40,57%
Random over most serendipitous (5% threshold)	42,57%
Random over most serendipitous (10% threshold)	46,67%
Random over most serendipitous (15% threshold)	48,67%

The results present a trend: with a larger threshold of randomness there are more good ratings. This could be interpreted as follows: the randomness of the selection of serendipitous item helps improving the ratings. So the best function would be the one with a more wide threshold of randomness. But, as the average c^+ ratings increase and better ratings means more similar items, we can hypothesize that suggested items are more semantically near to user tastes and knowledge so it is less probable that they are unknown. In this case the best function would be the most serendipitous one.

6. Conclusions and Future work

This paper reports a first effort to apply some ideas about serendipity to information retrieval and information filtering systems, especially in recommenders.

As future work, we expect to carry out more extensive experimentation with more users and wider item collections. We plan also to gather user feedback and feeling by questionnaires focused on qualitative evaluation of the recommendations and the idea of getting suggestions that should surprise them. That is really important for the need to understand the effectiveness of the module in finding unknown items rather the ones that result best rated. Experimentation

with users with different cultural levels and with different information seeking tasks are also important to find out which kind of user would like most serendipitous recommendations and to whom they are more useful.

We expect also to implement the other suggestions given by Toms [17] and to develop further the heuristic proposed (maybe padding a parameter factor that multiplies the probabilities in order to balance better between categories) or also introduce new heuristics and make an experimental comparison.

Acknowledgements

This research was partially funded by MIUR (Ministero dell'Università e della Ricerca) under the contract Legge 297/99, Prot.691 CHAT "Cultural Heritage fruition & e-Learning applications of new Advanced (multimodal) Technologies" (2006-08).

7. References

[1] P. Basile, M. Degemmis, A. L. Gentile, P. Lops, and G. Semeraro, "UNIBA: JIGSAW algorithm for Word Sense Disambiguation", Proc. of *4th ACL 2007 International Worshop on Semantic Evaluations (SemEval-2007)*, 2007.

[2] D. Billsus and M. J. Pazzani, "User Modeling for Adaptive News Access", *User Modeling & User-Adapted Interaction,* 10(2-3), pp. 147-180, 2000.

[3] R. Burke, "Hybrid Recommender Systems: Survey and Experiments", *User Modeling and User-Adapted Interaction,* 12(4), pp. 331-370, 2002.

[4] J. Campos and A. de Figueiredo, "Searching the Unsearchable: Inducing Serendipitous Insights", 2001.

[5] J. Campos and A. de Figueiredo, "The Serendipity Equations", Proc., 2001, pp. 121-124.

[6] E. De Bono, "Lateral Thinking", *Penguin Books*, 1990.

[7] M. Degemmis, P. Lops, and G. Semeraro, "A content-collaborative recommender that exploits WordNet-based user profiles for neighborhood formation", *User Modeling and User-Adapted Interaction: The Journal of Personalization Research (UMUAI),* 17(3), pp. 217-255, 2007.

[8] A. Foster and N. Ford, "Serendipity and information seeking: an empirical study", *Journal of Documentation,* 59(3), pp. 321-340, 2003.

[9] T. Gup, "Technology and the end of serendipity", *The Chronicle of Higher Education,* 44, p. A52, November 21, 1997.

[10] J. Herlocker, J. Konstan, L. Terveen, and J. Riedl, "Evaluating collaborative filtering recommender systems", *ACM Trans. Inf. Syst.,* 22(1), pp. 5-53, 2004.

[11] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", Proc. of *IJCAI*, 1995, pp. 1137-1145.

[12] S. M. McNee, J. Riedl, and J. A. Konstan, "Accurate is not always good: How Accuracy Metrics have hurt Recommender Systems", Proc. of *Extended Abstracts of the 2006 ACM Conference on Human Factors in Computing Systems (CHI 2006)*, 2006.

[13] G. A. Miller, "WordNet: a lexical database for English", *Communications of the ACM,* 38(11), pp. 39-41, 1995.

[14] D. Mladenic, "Text-learning and related intelligent agents: a survey", *IEEE Intelligent Systems,* 14(4), pp. 44-54, 1999.

[15] R. M. Roberts, "Serendipity: Accidental Discoveries in Science", *John Wiley & Sons, Inc*, 1989.

[16] G. Semeraro, M. Degemmis, P. Lops, and P. Basile, "Combining Learning and Word Sense Disambiguation for Intelligent User Profiling", Proc. of *IJCAI-07, 20th International Joint Conference on Artificial Intelligence*, Hyderabad, India, 2007, pp. 2856-2861.

[17] E. Toms, "Serendipitous Information Retrieval", Proc. of *DELOS Workshop: Information Seeking, Searching and Querying in Digital Libraries*, 2000.

[18] P. van Andel, "Anatomy of the Unsought Finding. Serendipity: Orgin, History, Domains, Traditions, Appearances, Patterns and Programmability", *British Journal Philosophy Science,* 45(2), pp. 631-648, 1994.