Review of Content-based Recommendation System

Prajakta A. Holey[1], Dr. S.S.Prabhune[2]
[1]M.E., C.E. Dept. of Comp. Sci. &Engg.
ShriSantGajananMaharaj College of Engineering, Shegaon, India.
[2]Professor& Head, Dept. of Information Technology,
ShriSantGajananMaharaj College of Engineering, Shegaon, India.

Abstract-- **Content-basedrecommendation systems try to recommend items similar to those a given user hasliked in the past.Recommender systems guides user in a personalizedway to interesting objects in a large space of possible options. Indeed, the basic process performed by a content-based recommenderconsists in matching up the attributes of a user profile in which preferencesand interests are stored, with the attributes of a content object (item), in order torecommend to the user new interesting items. It provides an overview ofcontent-based recommender systems, with the aim of imposing a degree of order onthe diversity of the different aspects involved in their design and implementation.**

Keywords:Recommendation, aggregation, utility matrix.

I. INTRODUCTION

The abundance of information available on the Web and in Digital Libraries, in combination with their dynamic and heterogeneous nature, has determined a rapidlyincreasing difficulty in finding what we want when we need it and in a manner whichbest meets our requirements. As a consequence, the role of user modeling and personalized information accessis becoming crucial: users need a personalized support in sifting through largeamounts of available information, according to their interests and tastes.Many information sources embody recommender systems as a way of personalizingtheir content for users [3]. Recommender systems have the effect of guidingusers in a personalized way to interesting or useful objects in a large space of possible options. Recommendation algorithms use input about a customer's intereststo generate a list of recommended items. At Amazon.com, recommendation algorithmsare used to personalize the online store for each customer, for example showing programming titles to a software engineer and baby toys to a new mother [4].The problem of recommending items has been studied extensively, and two mainparadigms have emerged. *Content-based* recommendation systems try to recommenditems similar to those a given user has liked in the past, whereas systemsdesigned according to the *collaborative* recommendation paradigm identify userswhose preferences are similar to those of the given user and recommend items they have liked [2].

Here, a comprehensive and systematic study of content-based recommendersystems is carried out. The intention is twofold:

• to provide an overview of state-of-the-art systems, by highlighting the techniqueswhich revealed the most effective, and the application domains in which they have adopted.

• to present trends and directions for future research which might lead towards thenext generation of content-based recommender systems.

The application of content-based recommendation system is 'Admission Recommendation System'for important role in Education System which allot the best colleges to the student according to their merit score for perusing higher education in reputed institutes e.g. (B.E./M.E) courses based on the cutoff marks at institute level in which they are eligible.Relevance Feedback model will be used with information retrival system in education system.

ISSN: 2278 – 7798

II. MECHANISM

Basics of Content-based Recommender Systems

Systems implementing a content-based recommendation approach analyze a set ofdocuments and/or descriptions of items previously rated by a user, and build a modelor profile of user interests based on the features of the objects rated by that user.The profile is a structured representation of user interests. The recommendation process basically consists in matchingup the attributes of the user profile against the attributes of a content object. The resultis a relevance judgment that represents the user's level of interest in that object.If a profile accurately reflects user preferences, it is of tremendous advantage forthe effectiveness of an information access process. For instance, it could be used tofilter search results by deciding whether a user is interested in a specific Web pageor not and, in the negative case, preventing it from being displayed.

A. A High Level Architecture of Content-based Systems

Content-based Information Filtering (IF) systems need proper techniques for representingthe items and producing the user profile, and some strategies for comparingthe user profile with the item representation. The high level architecture of a content-basedrecommender system is depicted in Figure (a). The recommendation processis performed in three steps, each of which is handled by a separate component:

• *Content Analyzer* –The main responsibility of the component is to represent the content of items(e.g. documents, Web pages, news, product descriptions, etc.) coming from informationsources. This representation is the input to the Profile Learner and Filtering Component.

• *Profile Learner* – This module collects data representative of the user preferencesand tries to generalize this data, in order to construct the user profile. Usually,the generalization strategy is realized through machine learning techniques i.e. user interests starting from items liked ordisliked in the past.

• *Filtering Component* – This module exploits the user profile to suggest relevantitems by matching the profile representation against that of items to berecommended.

Typically, it is possible to distinguish between two kinds of relevance feedback:positive information (inferring features liked by the user) and negative information(i.e., inferring features the user is not interested in).Two different techniques can be adopted for recording user's feedback.

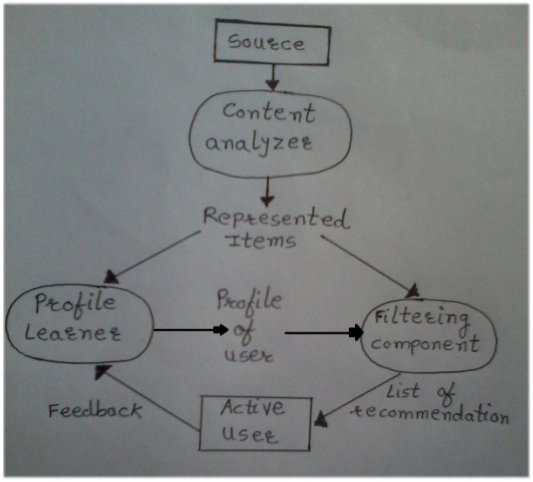

Fig. (a): Architecture of a Content-based Recommender

When asystem requires the user to explicitly evaluate items, this technique is usually referredto as "explicit feedback"; the other technique, called "implicit feedback", does not require any *active* user involvement, in the sense that feedback is derivedfrom monitoring and analyzing user's activities.Explicit evaluations indicate how relevant or interesting an item is to the user. There are three main approaches to get explicit relevance feedback.

• *like/dislike*– items are classified as "relevant" or "not relevant" by adopting asimple binary rating scale.

• *ratings*– a discrete numeric scale is usually adopted to judge items. Alternatively, symbolic ratings are mapped to a numeric scale, such as inSyskill&Webert [6], where users have the possibility of rating a Web page as*hot*, *lukewarm*, or *cold*;

• *text comments* – Comments about a single item are collected and presented to theusers as a means of facilitating the decision-making process.

ISSN: 2278 – 7798

The literature proposes advanced techniques from the affective computingresearch area to make content-based recommenders able to automaticallyperform this kind of analysis.Explicit feedback has the advantage of simplicity. Implicit feedback methods are based on assigning a relevance score to specific user actions on an item, such assaving, discarding, printing, bookmarking, etc.

B. Advantages and Drawbacks of Content-based Filtering

The adoption of the content-based recommendation paradigm has several advantageswhen compared to the collaborative one:

• *User Independence* - exploit solely ratingsprovided by the active user to build her own profile.

• *Transparency* - Explanations on how the recommender system works can beprovided by explicitly listing content features or descriptions that caused an itemto occur in the list of recommendations.

• *New item* - Content-based recommenders are capable of recommending itemsnot yet rated by any user.

•*Limited Content Analysis* - Content-based techniques have a natural limitin the number and type of features that are associated.

•*Over-Specialization* - Content-based recommenders have no inherent method for finding something unexpected.

• *New User* - Enough ratings have to be collected before a content-based recommendersystem can really understand user preferences and provide accuraterecommendations

III. METHODS

State of the Art of Content-based Recommender Systems

As the name implies, content-based filtering exploits the content of data items topredict its relevance based on the user's profile. Research on content-based recommendersystems takes place at the intersection of many computer science topics,especially Information Retrieval [5] and Artificial Intelligence.

From Information Retrieval (IR), research on recommendation technologies derivesthe vision that users searching for recommendations are engaged in an informationseeking process. In IR systems the user expresses a one-off information needby giving a query (usually a list of keywords), while in IF systems the informationneed of the user is represented by her own profile. Items to be recommended can be very different depending on the number and types of attributes used to describethem. Each item can be described through the same small number of attributes withknown set of values, but this is not appropriate for items, such as Web pages, news,emails or documents, described through unstructured text. From an Artificial Intelligence perspective, the recommendation task can be castas a learning problem that exploits past knowledge about users. At their simplest,user profiles are in the form of user-specified keywords or rules, and reflect the longterminterests of the user.

A. Item Representation

Items that can be recommended to the user are represented by a set of features,also called *attributes* or *properties*. For example, in a movie recommendation application,features adopted to describe a movie are: actors, directors, genres, subject matter, . . .)[7].

In most content-based filtering systems, item descriptions are textual featuresextracted from Web pages, emails, news articles or product descriptions. Unlikestructured data, there are no attributes with well-defined values. String matching suffers from problems of:

• POLYSEMY, the presence of multiple meanings for one word.

• SYNONYMY, multiple words with the same meaning.

The result is that, due to synonymy, relevant information can be missed if the profile does not contain the exact keywords in the documents while, due to polysemy, wrong documents could be deemed relevant. Items are represented as follows:

a.Keyword-based Vector Space Model

VSM is a spatial representation of text documents. In that model, each document is represented by a vector in a*n*-dimensional space, where each dimension corresponds to a

ISSN: 2278 – 7798

term from the overall vocabulary of a given document collection.

b. Review of Keyword-based Systems

Several keyword-based recommender systems have been developed in a relatively short time, and it is possible to find them in various fields of applications, such as news, music, e-commerce, movies, etc. Each domain presents different problems, that require different solutions.

c. Semantic Analysis by using Ontologies

Semantic analysis allows learning more accurate profiles that contain referencesto concepts defined in external knowledge bases. The main motivation for this approachis the challenge of providing a recommender system with the cultural andlinguistic background knowledge which characterizes the ability of interpreting naturallanguage documents and reasoning on their content.

d. Semantic Analysis by using Encyclopedic Knowledge Sources

Common-sense and domain-specific knowledge may be useful to improve the effectivenessof natural language processing techniques by generating more informativefeatures than the mere bag of words. The process of learning user profiles could benefitfrom the *infusion* of exogenous knowledge (externally supplied), with respect tothe classical use of endogenous knowledge (extracted from the documents themselves).

B. Methods for Learning User Profiles

The problem of learning user profiles can be cast as a binary text categorizationtask: each document has to be classified as interesting or not with respect to the user preferences. Therefore, the set of categories is $C = \{c+, c-\}$, where $c+$ is the positive class (user-likes) and $c-$ the negative one (user-dislikes).

a. Probabilistic Methods and Na¨ıve Bayes

Na¨ıve Bayes is a probabilistic approach to inductive learning, and belongs to the general class of Bayesian classifiers. These approaches generate a probabilistic model based on previously observed data.

b. Relevance Feedback and Rocchio's Algorithm

Relevance feedback is a technique adopted in Information Retrieval that helps users to incrementally refine queries based on previous search results. It consists of the users feeding back into the system decisions on the relevance of retrieved documents with respect to their information needs. Relevance feedback and its adaptation to text categorization, the well-knownRocchio'sformula , are commonly adopted by content-based recommender systems. The general principle is to allow users to rate documents suggested by the recommender system with respect to their information need.
The Rocchio's method is used for inducing linear, profile-style classifiers. This algorithm represents documents as vectors, so that documents with similar content have similar vectors.

C. Other Methods

Other learning algorithms have been used in content-based recommendation systems. Decision trees, Decision rule classifiers, nearest neighbor algorithms are some other methods usedin content-based recommendation system.

Decision trees are trees in which internal nodes are labeled by terms, branches departing from them are labeled by tests on the weight that the term has in the test document, and leaves are labeled by categories. Decision trees are used in the *Syskill&Webert*recommender system.

Decision rule classifiers are similar to decision trees, because they operate in a similar way to the recursive data partitioning approach. An advantage of rule learners is that they tend to generate more compact classifiers than decision trees learners.

Nearest neighbor algorithms, also called lazy learners, simply store training data in memory, and classify a new unseen item by comparing it to all stored items by using a similarity function. *Daily Learner* and *Quickstep* use the nearest neighbor algorithm.

IV. TRENDS AND FUTURE RESEARCH

A. The Role of User Generated Content in the Recommendation Process

ISSN: 2278 – 7798

Web 2.0 is a term describing the trend in the use of World Wide Web technologythat aims at promoting information sharing and collaboration among users. Accordingto Tim O'Reilly8, the term "Web 2.0" means putting the user in the center,designing software that critically depends on its users since the content, as in Flickr,Wikipedia, Del.icio.us, or YouTube, is contributed by thousands or millions of users.That is why Web 2.0 is also called the "participative Web". O'Reilly9 also definedWeb 2.0 as "the design of systems that get better the more people use them".

One of the forms of User Generated Content (UGC) that has drawn more attentionfrom the research community is *folksonomy*. The freely chosen keywords according to interest are called *tags*.

Folksonomies provide new opportunities and challenges in the field of recommendersystems.

a. Social Tagging Recommender Systems

Several methods have been proposed for taking into account user tagging activitywithin content-based recommender systems.

The user profile is represented in the form of a tag vector, with eachelement indicating the number of times a tag has been assigned to a document by thatuser. The matching of profiles to information sources is achieved byusing simple string matching.

B. Beyond Over-specialization: Serendipity

Content-based systems suffer from over-specialization,since they recommend only items similar to those already rated by users. One possiblesolution to address this problem is the introduction of some randomness. For example,the use of genetic algorithms has been proposed in the context of information filtering. In certain cases, items should not be recommended if they are too similar to something the user has already seen, such as a different news articledescribing the same event. In summary, the *diversity* of recommendations isoften a desirable feature in recommender system.

It is useful to make a clear distinction between *novelty* and *serendipity*. Novelty occurs when the system suggests to the user anunknown item that she might have autonomously discovered. A serendipitous recommendationhelps the user to find a surprisingly interesting item that she mightnot have otherwise discovered (or it would have been really hard to discover). The example of the difference between novelty and serendipity, consider a recommendation system that simply recommends movies that were directed bythe user's favorite director. If the system recommends a movie that the user wasnot aware of, the movie will be novel, but probably not serendipitous. On the otherhand, a recommender that suggests a movie by a new director is more likely to provide serendipitous recommendations. Recommendations that are serendipitous areby definition also novel.

V. CONCLUSIONS

In this content-based recommender systems, is overviewed and provided by the most important characterizing systems. Although there is a bunch of recommender systems in different domains,they share in common a means for representing items to be recommended and user profiles. Here the main issues related to the representation of items,starting from simple techniques for representing structured data, to more complextechniques coming from the Information Retrieval research area for unstructureddata are discussed. The main content recommender systems developed in the last 15years, by highlighting the reasons for which a more complex "semantic analysis"of content is needed in order to go beyond the syntactic evidence of user interestsprovided by keywords are analyzed. A review of the main strategies (and systems) adopted tointroduce some semantics in the recommendation process is carried out, by providingevidence of the leading role of linguistic knowledge, even if a more specificknowledge is mandatory for a deeper understanding and contextualization of theuser interests in different application domains.

ISSN: 2278 – 7798

The issues presented here will contribute to stimulate their search community about the next generation of content-based recommendationtechnologies.

REFERENCES

[1] Pasquale Lops, Marco de Gemmis and Giovanni Semeraro:ch3-Content-based Recommender Systems: State of the Art and Trends © Springer Science+Business Media, pp-73-105, LLC (2011).

[2] Mining the Massive Databases, ch.9-Recommendation systems.,AnandRajaraman, J.D.Ullman, Cambridge University Press, 2013.

[3] Resnick, P., Varian, H.: Recommender Systems. Communications ofthe ACM 40(3), 56-58, (1997).

[4] Linden, G., Smith, B., York, J.: Amazon.com Recommendations: Item-to-Item Collaborative Filtering. IEEE Internet Computing 7(1), 76–80 (2003).

[5] Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley (1999).

[6] Pazzani, M.J., Muramatsu, J., Billsus, D.: Syskill and Webert: Identifying Interesting WebSites. In: Proceedings of the ThirteenthNational Conference on Artificial Intelligence andthe Eighth Innovative Applications of Artificial Intelligence Conference, pp. 54–61. AAAIPress / MIT Press, Menlo Park (1996).

[7] Pazzani, M.J., Billsus, D.: Content-Based Recommendation Systems. In: P. Brusilovsky, A. Kobsa, W. Nejdl (eds.) The Adaptive Web, *Lecture Notes in Computer Science*, vol. 4321, pp. 325–341 (2007). ISBN 978-3-540-72078-2.

ISSN: 2278 – 7798