Content-based Recommender System using Social Networks for Cold-start Users

Alan V. Prando[1], Felipe G. Contratres[2], Solange N. A. Souza[2] and Luiz S. de Souza[3]
[1]*Instituto de Pesquisas Tecnológicas, São Paulo, Brazil*
[2]*Universidade de São Paulo, São Paulo, Brazil*
[3]*Faculdade de Tecnologia, São Paulo, Brazil*

Keywords: Recommender System, Social Networks, Cold-Start, Content-based Recommendations.

Abstract: Recommender systems have been widely applied to e-commerce to help customers find products to purchase. Cold-start is characterized by the incapability of recommending due to the lack of enough ratings. In fact, solutions for the cold-start problem have been proposed for different contexts, but the problem is still unsolved. This paper presents a RS for new e-commerce users by using only their interactions in social networks to understand their preferences. The proposed Recommender System (RS) applies a content-based approach and improves the experience of new users by recommending specific products in a preferred identified user category by analysing their data from the social network. Therefore, it combines three social network elements: (1) direct user posts (e.g.: "tweets" from Twitter and "posts" from Facebook), (2) content "likes" (e.g.: option "like" on a "post" or "tweet" posted by another user), and (3) page "likes" (e.g.: option "like" on a Facebook page). The proposed RS was tested for a retail e-commerce, which usually not only has a large range of categories of products, but also has products within these categories. The difficulty in predicting a product increases sharply with a greater number of categories and products. According to the experiment conducted, the proposed RS demonstrated to be a reasonable alternative to cold-start, i.e., for users accessing e-commerce for the very first time.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

Data volume has grown continuously over the last years. Companies started to store all sorts of data, from server logs to any useful information with business value (Landim et al. 2013) (Fan & Bifet 2013) (Singh & Singh 2012). As a consequence, our ability to collect data from various applications in different formats has been increasing drastically. This data – and the technologies capable to deal with them - is called Big Data (Madden 2012), which has truly affected current business. A few years ago, an enterprise used to store a restricted subset of data containing only core information. In contrast, nowadays an enterprise can leverage all the information available with large data to gain insights and make better decisions, having an advantage over the market competitors (Singh & Singh 2012).

Recommender Systems are a prime example of the mainstream applicability of Big Data. With applications such as e-commerce and music/video streaming, services use recommender systems techniques to mine and to process large volumes of data to better match the needs of their users in a personalized fashion (Fan & Bifet 2013). Above all, recommender systems have emerged as a way to help users in their decision-making process because they suggest the most suitable items to a particular user (Adomavicius & Tuzhilin 2005). Recommender system (RS) encompasses personalized algorithms that use machine learning (ML) and data mining techniques to identify the preference of each user individually (Yang et al. 2011) (Ricci et al. 2011). It is common for e-commerce to employ recommender systems as a differential (Linden et al. 2003).

However, sometimes there is a lack of data and the RS cannot predict user preference. This occurs, for example, when a new user registers in the e-commerce and the system may not have enough information about him/her for recommending. This problem is known as cold-start (Ricci et al. 2011) (Adomavicius & Tuzhilin 2005) (Sun et al. 2015) (Fernández-Tobías et al. 2016).

Prando A., Contratres F., Souza S. and de Souza L.
Content-based Recommender System using Social Networks for Cold-start Users.
DOI: 10.5220/0006496301810189
In *Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KDIR 2017)*, pages 181-189
ISBN: 978-989-758-271-4
Copyright © 2017 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

This paper introduces a RS for e-commerce by using only social network data to improve recommendations for the cold-start problem. More specifically, predicting products for the user to buy is solely made by using posts from users' social networks. As a result, the RS uses techniques that correlate e-commerce products to the elements of social networking. Thus, even if there are no explicit or implicit ratings the data extracted from the social networks may be sufficient to determine user preference.

This paper is organized as follows. Section 2 presents the main aspects of previous work on Recommender Systems (RS). Section 3 introduces the recommender system definition, besides including the main characteristics of social networks. Section 4 details the proposed RS. Section 5 presents the experiment implemented to evaluate the proposed RS and its results. Finally, section 6 concludes it and suggests future work.

2 RELATED WORK

A mix of algorithms to obtain implicit and explicit evaluations or to infer user action have been suggested in previous work (Schein et al. 2002) (Burke 2007) (Konstan & A. 2004) to solve the cold-start problem. Despite the efforts made, the lack of information and ratings are still a problem (Ricci et al. 2011) (Fernández-Tobías et al. 2016).

Fernándes-Tobías et al. (Fernández-Tobías et al. 2016) proposes colllaborative-filtering (CF) techniques (Adomavicius & Tuzhilin 2005) (Burke 2007) (Ricci et al. 2011) (Kaššák et al. 2016) that include personality information and cross domain information to recommend items to new users. Their proposal assumes that, in certain domains, users with simliar personalities tend to have similar preferences. They employ the Big Five test (Quirino, Mals, Groterhorst, De Souza, et al. 2015), a personality test to outline the profile of a person, to define users' personality trait and the system thus has information about the users' pesonalities. Moreover, based on personality traits, they propose the CF technique to extract user preferences in the auxiliar domains and to apply this information to recommend itens in a target domain to a new user. Good results were derived with this personality-aware method. However, new users have to take the Big Five test before interacting with the recommender system, which, in a real case, may discourage them to keep using e-commerce.

Considering the lack of data, social networks appear as natural sources of data for recommendation and a feasible solution for cold-start. In recent years, different works have employed data from social network to recommend products and as a solution for cold-start.

Ma et al. (Ma et al. 2011) proposes a method that forms a social composition, according to the common interests among the friends of a particular user. Following the same idea, a recommendation is made for a particular user based on the social influence of both close and distant friends (He & Jianming 2010). Oliveira et al. (Oliveira et al. 2012) built a RS based on trust among friends, called trust-aware recommender systems. In the same fashion, but considering the cold-start problem, Caron, and Bhagat (Caron & Bhagat 2013) use the information of a new user's friends in e-commerce and propose a learning model of user preference in the social network. Lalwani et al. (Lalwani et al. 2015) assume users that are part of a community have similar preferences, or are influenced by it. They utilize social interactions (friend connections) to detect Facebook communities and to simulate cold-start scenarios, recommending items already ranked by users in the community to the user without an assessment history.

Maniktala et al. (Maniktala et al. 2015) also assume users prefer items already acquired by friends, but the friends' influence depends on how strong the friendship relation is. They propose techniques to classify social relationships in strong or weak and recommend items acquired by users with strong friend connection to a new user. However, they consider cold-start for a user that has consumed up to 5 items, which can be considered a moderate cold-start once there is some information about the user.

Felicio et al. (Felicio et al. 2016) use models already built in systems for users to suggest items to a new user. They built a personalized model for a cold-start user by selecting prediction models from a set of strongly linked users. They handle several social network connection weight metrics to classify links among users.

Amatriain (Amatriain & Xavier 2013) presents the main algorithms used in RS. Zhang and Pennacchiotti (Zhang & Pennacchiotti 2013b) (Zhang & Pennacchiotti 2013a) employ data mining techniques to extract information from social networks to find a users' preferred categories in e-commerce. In their experiment, classes of goods from Ebay, which were previously ranked by the user, are associated with a class of user-liked pages

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

(user explicitly indicates the interest in a determined Facebook page). In this experiment, specific products in categories are not considered.

The data-mining field has considerably advanced in recent years due to technological advances providing the processing and storage of a large volume and variety of data. In particular, social networks are considered essential to this change, driving the creation of such data, which is generated by different users. Thus, individual efforts regarding techniques for extracting information on specific data types, such as text mining, become relevant for achieving results (Aggarwal & Zhai 2012) (Hu & Liu 2012) (Xia et al. 2016).

Hu and Liu (Hu & Liu 2012) introduce a text-mining model for social network divided into three stages: (1) Text process: incoming documents are manipulated to achieve appropriate representation – stop-word removal and stemming are some of the techniques employed. (2) Text representation: documents are transformed into sparse numeric vectors (bag-of-word, BOW, or vector space model, VSM) – an algebraic model for representing text document as a vector of identifiers (ex. index terms), in which the relevance ranking of documents may be calculated. Term Frequency-Inverse Document Frequency (TF-IDF) is a common technique to weight each term in a document, scoring the importance of the words in a document based on how frequently they appear across multiple documents. (3) Knowledge discovery: having documents represented by a vector - ML algorithms can be applied to find document similarities.

Similarly to Zhang and Pennacchiotti (Zhang & Pennacchiotti 2013b) (Zhang & Pennacchiotti 2013a), our proposed RS uses data extracted from users' social networks, but in our case, no previously ranked purchase or evaluations given by users are employed. In other words, our experiment considers that users are really new in the e-commerce and no information about them exists, which is classified as an extreme cold-start scenario (Fernández-Tobías et al. 2016). Moreover, differently from most of the previous works, our RS does not use users' friends' data. Therefore, the proposed RS contributes by showing that a content-based approach (Adomavicius & Tuzhilin 2005) (Burke 2007) (Ricci et al. 2011) (Kaššák et al. 2016) may also be employed, yielding good results. Our RS investigates whether data posted by the user himself may be used as a complement to other techniques to make more assertive recommendations.

Despite the variety of data in social networks, text is clearly predominant. Therefore, the proposed RS applies text mining techniques and ML algorithms to perform recommendations by correlating user social interactions with e-commerce products, characterizing a content-based recommendation.

3 RECOMMENDER SYSTEM AND PROBLEM STATEMENT

Formally, recommendation problems can be formulated as: C is the set of all users and S is the set of items that might be recommended. Utility function u measures how useful item s is to a particular user c, e.g., u: C x S → R, where R is an ordered set. Hence, for each user c ∈ C, an item s ∈ S that maximizes the utility to user c is searched; this is represented by equation 1 (Adomavicius & Tuzhilin 2005):

$$\forall c \in C, s_c = \arg\max_{s \in S}(u, s) \quad (1)$$

In RS, the utility of an item is represented by a rating, which points out the interest of a user in a particular item. Each user in C has many attributes such as age, gender, marital status, etc. Similarly, each item in S is defined by its characteristics such as description, specifications, etc. Therefore, the proposed RS is content-based, since for item s and user c, the function u(c, s) estimates the rating given by user c to item s, based on utilities u(c, s_i) assigned by user c to items $s_i \in S$, which are similar to item s.

Accordingly, to solve the cold-start problem, the proposed RS uses data mined from users social networks (e.g., recent social data, such as tweets and posts) as implicit ratings to calculate u(c, s).

3.1 Social Networks

The last decade represented a revolution in the way society interacts due to the intense use of social networks. This interaction has become a powerful tool for analysis and knowledge discovery regarding its users and the way in which they communicate with each other. Social networks are a type of service offered by the web that allows its users to exchange information (Quirino, Mals, Groterhorst, de Souza, et al. 2015) (Albalooshi et al. 2012). Recently, social networks such as Facebook, Twitter, LinkedIn and Foursquare became extremely popular in the whole world. The number of Facebook users increased 20 times in the 2008 - 2012 period (Long Jin et al. 2013). In additon, a large number of different kind of devices such as desktops, notebooks

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

mobile, smartphones and tablets have been facilitating the use of social networks (Long Jin et al. 2013).

Social networks intend to provide security an privacy for its users (Joshi & Kuo 2011). Each social network has different approaches to deal with user privacy. For example, Twitter exposes all posts published, whereas LinkedIn restricts data accessed according to the type of user account and relationship. Facebook makes available an Application Programming Interface (API) to other applications to connect to it and the number of systems, which allow their users to register or to connect to them by using a Facebook account, have increased. If an user opts for it, the application recovers the user's personal data from Facebook, such as name, date of birth, etc. In general, only personal data is transmitted to the applications from Facebook, but if the user gives permission, other information can be accessed.

The social networks employed as data source to the proposed RS was Facebook and Twitter because of their popularity, considering the number of users (Quirino, Mals, Groterhorst, de Souza, et al. 2015) (Derczynski et al. 2015). Besides, the access to Twitter is free, facilitating the access of user data, although only the last 200 tweet users are acessible. In case of Facebook, an a application was created to allow accessing the users data.

4 CHARACTERISTICS OF PROPOSED RECOMMENDER SYSTEMS

4.1 Recommendation Process

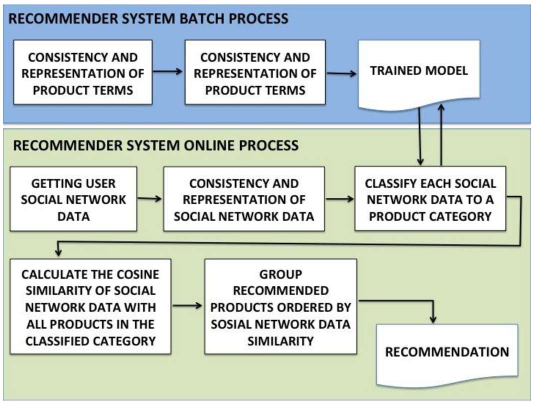

RECOMMENDER SYSTEM BATCH PROCESS
CONSISTENCY AND REPRESENTATION OF PRODUCT TERMS
CONSISTENCY AND REPRESENTATION OF PRODUCT TERMS
TRAINED MODEL
RECOMMENDER SYSTEM ONLINE PROCESS
GETTING USER SOCIAL NETWORK DATA
CONSISTENCY AND REPRESENTATION OF SOCIAL NETWORK DATA
CLASSIFY EACH SOCIAL NETWORK DATA TO A PRODUCT CATEGORY
CALCULATE THE COSINE SIMILARITY OF SOCIAL NETWORK DATA WITH ALL PRODUCTS IN THE CLASSIFIED CATEGORY
GROUP RECOMMENDED PRODUCTS ORDERED BY SOSIAL NETWORK DATA SIMILARITY
RECOMMENDATION

Figure 1: Flowchart of recommendation.

The recommendation process is divided into two main flows (Fig. 1): (1) RS batch process, which runs only once a day (or when a new product is included); (2) RS online process, which runs on each recommendation request. To our knowledge, this is the first study that creates a process (Fig. 1) able to recommend using only users' social network data to match products from real e-commerce employing content-based approach. Although the techniques employed have already been discussed in the Text Mining and Information Retrieval research areas, the combination of these techniques to achieve a content-based recommendation by solely using social networks data can be considered the main contribution of this work.

According to Fig. 1, the tasks performed are:

a) Consistency and representation of product terms and social network data:

First, the consistency step applies the stemming and stop word removal techniques (Hu & Liu 2012) to remove generic terms and normalizes both social media data and product data. Moreover, the representation step also calculates TF-IDF (eq. 2) to prepare the data for the next task.

TF-IDF is applied in batch and in an online process. More formally, in the batch process, TF-IDF (eq. 2) is represented by a set of product features S with a list of terms t representing each word of the feature.

$$tfidf(s,t) = \frac{occurProdT(s,t)}{totProdT(s)} \cdot log\frac{|S|}{|(S,t)|} \qquad (2)$$

Where:

- occurProdT(s,t) is the total of terms in product s.
- $totProdT(s)$ is the total of terms in product s.
- $|(S,t)|$ is the number of products with term t.
- $|S|$ is the total of products.

As a consequence, a Vector Space Model (VSM) represents all the products with terms weighted for each product. The VSM is used in the training task.

Similarly, in the online process, TF-IDF uses a set of social data (posts) P of user c, with a list of terms t representing each word of P (eq. 3).

$$tfidf(p,t) = \frac{occurSocialT(p,t)}{totSocialT(p)} \cdot log\frac{|P|}{|(P,t)|} \qquad (3)$$

Where:

- $occurSocialT(p,t)$ is the total of terms in social data p.
- $totSocialT(p)$ is the total of terms in social data p.
- $|(P,t)|$ is the number of social data with term t.
- $|P|$ is the total of products.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

Therefore, all social data are represented by a VSM allowing the computation of the similarity distance between social data and product.

b) Training of Classifier using product terms as input and category as classes:
The training phase starts as soon as RS is available. It receives the VSM of product features as input and the product categories as classes. Therefore, after training, the classifier is used to determine the category of products the social data belong to. When the training phase is completed, the RS is ready to receive requests from e-commerce. Each request uses the classifier in isolation with the VSM of social data to predict the preferred user category. Since each social datum derives a product category, a function named classrank (eq. 4) is employed for ordering the preferred classes of user c, with f representing a set of categories.

$$classrank(c,f) = \frac{social(c,f)}{\sum_{f_i \in F} social(c,f_j)} \quad (4)$$

Where: $social(c,f)$ is the total of social data from user c classified in class f, establishing the order, according to eq. 5:

$$f_i > f_j \Leftrightarrow classrank(c,f_i) > classrank(c,f_j) \quad (5)$$

In summary, eq. 5 orders the categories by its importance, which is measured by the number of times a social data is related to.

c) Acquisition of user social networks data
Not all data from social networks are available or can be used to determine user preferences. Thus, the data from social network used by the classifier in the proposed RS are:

- Likes on pages (from Facebook). Social networks establish pre-set categories that can be related to e-commerce categories. Zhang and M. Pennacchiotti (Zhang & Pennacchiotti 2013b) also used this approach.
- Likes on contents (from Facebook and Twitter) are used in the classification and product similarity phases. The content can have a text format employed by users to express their opinion to friends (Albalooshi et al. 2012) (Long Jin et al. 2013).
- Post (from Facebook) and “Tweets” (from Twitter) are used in the classification and product similarity phases. Both are texts published by users to express their opinion to friends (Albalooshi et al. 2012) (Long Jin et al. 2013).
- Shares (from Facebook and Twitter), which are also utilized by users to express their opinion to friends (Albalooshi et al. 2012) (Long Jin et al. 2013) (Fersini et al. 2016) (Saif et al. 2016).

The intention is to perform a more personalized recommendation than the one proposed by Zhang (Zhang & Pennacchiotti 2013b). For instance, a RS for sportswear e-commerce, in the case of a user liking a football page and publishing a post with the word “Barcelona", a list called "Football" with products related to the Barcelona Club will be recommended.

In the same example, Zhang recommends (Zhang & Pennacchiotti 2013b) a list called "football", but no products related to Barcelona will necessarily be recommended.

d) Similarity between social network data and products in the classified category:
The preferred categories of users are obtained from the data extracted from their social network. However, since each category has a large dataset of products, it is hard to find what product is more similar to the data produced by the user social data in a category. In order to provide the best recommendation, the cosine similarity (Adomavicius & Tuzhilin 2005) (Amatriain & Xavier 2013) (Amatriain 2013) is applied to each product (between each vector P - representing a particular product, and vector S - representing the user social data recovered), yielding an ordered list of the products most similar to the user social data.

4.2 Supervised Machine Learning Algorithms Employed

The architectures employed is modular and extensible and is an extension of (Prando & Alves de Souza 2016). RS should be able to obtain user social data and to process the recommendation in a suitable time for e-commerce to use it, i.e., during the visit of the user to e-commerce (Amatriain & Xavier 2013). Big Data technologies are employed in proposed RS (Fan & Bifet 2013) (Rios & Diguez 2014), (Lin & Ryaboy 2013), which could be scaled to reach e-commerce response time requirements.

The proposed Recommender systems use Naïve Bayes, Decision Tree and SVM (Support Vector Machine), supervised ML algorithms, to classify products by their characteristics, representing the content-based recommendation approach (Rokach & Maimon 2007) (Joachims & Thorsten 2002) (Rennie et al. 2003).

The classifiers utilized in the experiment were dictated by the technologies employed, namely

Apache Spark and Apache Hadoop. Consequently, only Naïve Bayes and Decision Tree were used because SVM multiclass was not found in Apache Spark when the experiment was carried out.

In order to define the classifier to be employed, Naive Bayes or Decision tree, two variables were observed: (1) prediction response time, and (2) performance – returned success rates by the classifier for an 80/20 comparison, in which 80% of a product data is employed in the training phase and 20%, in the assessment phase. The categories of product tree are handled as classes for the classifiers. In Fig. 1, flows (1) and (2) of the proposed RS are strongly affected by the hierarchic level chosen.

Classifier performance in flow (1) is affected as follows: category in a higher level has more products in it, increasing the class features for training – contributing to achieving better performance (e.g. the "TV and Home Theatre" root category contains the whole TVs and home theatres while the "LED TV" subcategory only has LED TVs of the e-commerce). On the other hand, in flow (2), computational complexity and the recommendation time processing increase the higher the category level, because this implies more products in a class. Consequently, more calculation is conducted in the similarity cosine phase. The category of social data in the e-commerce context is predicted in the social data classification phase, Fig. 1 - flow (2). Then, in the next similarity cosine phase, this social data are compared with all the products of the category identified, allowing products more similar to social data to be recommended.

Table 1 shows the results of Naïve Bayes (NB) and Decision Tree (DT) classifiers evaluation, using two entries: (1) VSM using root categories as a class; (2) VSM using second level categories as a class. As a result, a more appropriate response time and performance to process a task for a recommendation during an interaction on an e-commerce were achieved with Naïve Bayes.

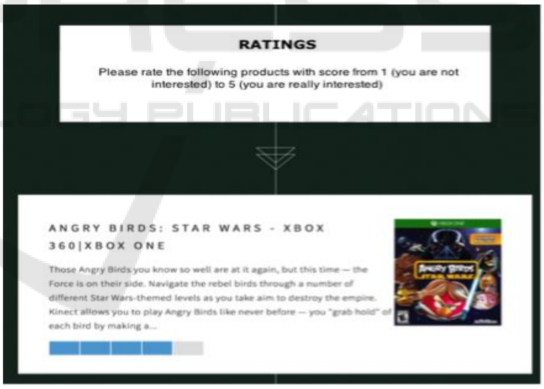
Table 1: Evaluation of Naïve Bayes and Decision Tree classifiers.

Algorithm	Prediction time (ms)	Performance 80/20
NB - root categ. (NBC)	51	0.96
NB – 2º level categ.	37	0.68
DT - root categ.	120	0.53
DT – 2º level categ.	124	0.09

In summary, the classification process can be detailed as follows: the classifier is trained using a VSM, which is calculated by employing TF-IDF of the description of each product, having classes represented by product categories. This way, the NBC classifier (Table 1) learns the terms of products distributing them into categories, allowing the classification of any other set of terms with similar representation. Hence, with the model trained, RS is ready to receive requisitions, which are social data retrieved and represented by VSM of terms using TF-IDF. Since each social datum has its own representation, it might be classified in a product category, concluding the classification process.

5 EXPERIMENT

In order to test the proposed RS and to show the results, an application was created to simulate a user's first time in e-commerce. This application was employed because of the available budget. The use of a real scenario will demand a more robust infrastructure than the one used. For the experiment, 1.449 products divided into 239 categories were used, arranged into 15 categories in the first level (root category) and 67 in the second level (daughters of root category). These products are a subset of the products extracted from the e-commerce BestBuy by using an API available on the web.



RATINGS

Please rate the following products with score from 1 (you are not interested) to 5 (you are really interested)

ANGRY BIRDS: STAR WARS - XBOX 360|XBOX ONE

Those Angry Birds you know so well are at it again, but this time — the Force is on their side. Navigate the rebel birds through a number of different Star Wars-themed levels as you take aim to destroy the empire. Kinect allows you to play Angry Birds like never before — you "grab hold" of each bird by making a...

Figure 2: Form: assessment of products.

The steps of the application created are:

1. The user must accept to participate in scientific research, having prerequisites: legal age and being registered in Facebook and/or in Twitter.
2. The user logs in by using Facebook or informing his/her Twitter account. Therefore, user data is obtained from the social networks Facebook and Twitter.
3. The recommendation flow starts and products are shown.
4. The user must rate the products recommended by RS, assigning a score of 1 (no interest) to 5 (high interest). The value attributed to the

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

recommendation by the user is employed as metrics in the RS assessment phase (Fig. 2).

5. User submits the evaluations.

The application was distributed to people selected in advance. The authors invited people (friends, colleagues, students at university) to take part in the experiment. To conduct the controlled acceptable experiment, a minimum of 68 individuals or participants is necessary, which allows estimates with an error-margin of 10% and confidence level of 90% (Krejcie & Morgan 1970).

Experiment data were organized in 4 text files: users, products, social_data and assessment. Table 2 shows file name, its structure (file fields) and an example of data for each field.

Table 2: Structure of files with data sample of experiment.

Users File	
name	user.txt
structure	User ID; gender; age; source
data sample	1; M; 28; Facebook
Products File	
name	product.txt
structure	product ID; name; description; category
data sample	1; iphone; apple iphone; smartphone;
Social Data File	
name	social_data.txt
structure	user ID; BOW of data
data sample	1; smartphones, nice; eat, past;
Assessment File	
name	Assessment.txt
structure	user ID; product ID; assessment;
data sample	1; 1; 3

5.1 Experimental Results and Assessment

Individuals that participated in the experiment did not receive any advertisement regarding its content or goal, and totalled 98 participants. In summary, the following results were achieved:

- It was impossible to generate recommendation for 16 participants due to the lack of social data or the NB classifier could not predict the class of the social data captured.
- 72 participants finished the whole process.
- 718 recommendations were generated and rated.

The result of cosine similarity yields values in the [-1, 1] interval. All the results below 0.02 were discarded, this range of values being considered to represent no similarities between users' social data or product features. Then, the distribution of cosine similarity results from 0.02 (smallest) to 0.7 (highest). 65% of the results are observed to be between 0.02 and 0.2.

Furthermore, the participants rated each recommendation given by the proposed RS (Fig. 2) with grades from 1 to 5. To evaluate performance, the root mean square error (RMSE) (Shani & Gunawardana 2011) is used to measure the difference between both ratings. The use of RMSE is very common in the field and it makes a good error metric for RS (Ricci et al. 2011) (Fernández-Tobías et al. 2016) (Amatriain 2013). To calculate the RMSE between users and RS ratings, firstly RS ratings need to be normalized because they generate a numerical value in the interval of [-1, 1], whereas user ratings attributes values in the interval of [1, 5]. Therefore, RS results were normalized to values in the range of [1, 5] (eq. 6). In equation 6, let P and S be the set of products and social data of users, respectively. Let p_{ij} be the product returned by RS to user $i, p_{ij} \in P$.

$$normalizeRating(i, p_j) = \frac{\cos(s_{i,j}, p_{i,j})}{\max\limits_{s \in S, p \in P} \cos(s, p)} * 5 \qquad (6)$$

Where: $normalizeRating(i, p_i)$ is the RS rating for social data of user i in product $p_j \in P$ and $s_{i,j}$ is the social data related to product $p_{i,j}$.

It is now possible to calculate the RMSE (eq. 7). Let $q_{p,u}, r_{p,u}$ be the grades attributed by user u and RS to product p, respectively.

$$RMSE = \sqrt{\frac{1}{NM} \sum_{p=1}^{N} \sum_{u=1}^{M} \left(q_{p,u} - r_{p,u}\right)^2} \qquad (7)$$

Where:

- N is the total recommendations made.
- M is the total of users.
- $1 \leq q_{p,u} \leq 5 \; and \; 1 \leq r_{p,u} \leq 5.$

Considering equation 7, the RMSE value can vary from 0 (best case, i.e., the grades given by the user and cosine similarity are much closer) to 4 (worst case, i.e., the grades given by user and cosine similarity are much more distant). Ranges are thus defined to interpret RMSE results, such as: [0, 1] – very good; (1, 2] - good; (2, 3] - bad; (3, 4] - very bad. Finally, the result for RMSE was 1.71. According to the ranges defined, the RS performed well and can be considered a good alternative to the cold-start.

Complementing the results from the RMSE, Table 3 shows the numbers of occurrences in which grades given to the product offered by the proposed

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

RS and users are equal. According to this result, it is possible to affirm that proposed RS has 40% accuracy. This reveals that user’s social data, i.e., data extracted from the social network of a user, who is receiving the recommendation, is an important source of data for recommendation processes, mainly for the cold-start problem, in which no data about the user exists. Besides, the experiment was conducted for a more generic e-commerce than those of movies or music, confirming the efficiency of the proposed RS.

Table 3: Correspondence between grades given by the proposed RS and attributed by the user.

Difference between grades attributed by the proposed RS and users	Number of occurrences	Percentage
0	288	40.11
> 0	430	59.89

6 CONCLUSIONS

To demonstrate how the proposed RS responds to the cold-start problem, an experiment that recommends products by reproducing the conditions of an e-commerce to a new user was built and distributed to a group of participants. As a result, a satisfactory RMSE of 1.71 was scored. Additionally, the results were complemented by counting the number of occurrences, in which the grades given by user and proposed RS coincided, showing that the proposed RS had 40% accuracy. It demonstrates that the proposal can be a good alternative to the cold-start problem. Besides, the proposed RS employed a content-based approach and the results showed that this technique is appropriate to cold-start and, if it is used as a complement to the collaborative-filtering approach, it may even further improve the results from a cold-start problem.

There are many future important challenges in RS using social networks that arise from the nature of personalization: discovering user preferences by processing a sparse, diverse and large dataset employing techniques that demand a high computational capacity. As continuity of this work is intend to: (i) use sentiment analysis technique to evaluate the sentiment represented in social data before use it in recommendation process (ii) encompass data from social network of friends network, i.e., not only use data from user social network, but also from his/her network of friends.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

ACKNOWLEDGEMENTS

Authors are grateful for the support given by São Paulo Research Foundation (FAPESP). Grant #2014/04851-8, São Paulo Research Foundation (FAPESP).

REFERENCES

Adomavicius, G. & Tuzhilin, A., 2005. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), pp.734–749.

Aggarwal, C.C. & Zhai, C., 2012. An Introduction to Text Mining. In *Mining Text Data*. Boston, MA: Springer US, pp. 1–10.

Albalooshi, N., Mavridis, N. & Al-Qirim, N., 2012. A survey on social networks and organization development. In *2012 International Conference on Collaboration Technologies and Systems (CTS)*. IEEE, pp. 539–545.

Amatriain, X., 2013. Big &amp; personal: data and models behind netflix recommendations. In *Proceedings of the 2nd International Workshop on Big Data, Streams and Heterogeneous Source Mining Algorithms, Systems, Programming Models and Applications - BigMine ’13*. New York, New York, USA: ACM Press, pp. 1–6.

Amatriain, X. & Xavier, 2013. Mining large streams of user data for personalized recommendations. *ACM SIGKDD Explorations Newsletter*, 14(2), p.37.

Burke, R., 2007. Hybrid Web Recommender Systems. In *The Adaptive Web*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 377–408.

Caron, S. & Bhagat, S., 2013. Mixing bandits. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis - SNAKDD ’13*. New York, New York, USA: ACM Press, pp. 1–9.

Derczynski, L. et al., 2015. Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2), pp.32–49.

Fan, W. & Bifet, A., 2013. Mining big data: current status, and forecast to the future. *ACM SIGKDD Explorations Newsletter*, 14(2), p.1.

Felicio, C.Z. et al., 2016. Preference-Like Score to Cope with Cold-Start User in Recommender Systems. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, pp. 62–69.

Fernández-Tobías, I. et al., 2016. Alleviating the new user problem in collaborative filtering by exploiting personality information. *User Modeling and User-Adapted Interaction*, 26(2), pp.1–35.

Fersini, E., Messina, E. & Pozzi, F.A., 2016. Expressive signals in social media languages to improve polarity detection. *Information Processing & Management*, 52(1), pp.20–35.

He & Jianming, 2010. *A social network-based recommender system*. University of California at Los Angeles.

Hu, X. & Liu, H., 2012. Text Analytics in Social Media. In *Mining Text Data*. Boston, MA: Springer US, pp. 385–414.

Joachims, T. & Thorsten, 2002. *Learning to classify text using support vector machines*, Kluwer Academic Publishers.

Joshi, P. & Kuo, C.-C.J., 2011. Security and privacy in online social networks: A survey. In *2011 IEEE International Conference on Multimedia and Expo*. IEEE, pp. 1–6.

Kaššák, O., Kompan, M. & Bieliková, M., 2016. Personalized hybrid recommendation for group of users: Top-N multimedia recommender. *Information Processing & Management*, 52(3), pp.459–477.

Konstan & A., J., 2004. Introduction to recommender systems J. A. Konstan, ed. *ACM Transactions on Information Systems*, 22(1), pp.1–4.

Krejcie, R. V. & Morgan, D.W., 1970. Determining Sample Size for Research Activities. *Educational and Psychological Measurement*, 30(3), pp.607–610.

Lalwani, D., Somayajulu, D.V.L.N. & Krishna, P.R., 2015. A community driven social recommendation system. In *2015 IEEE International Conference on Big Data (Big Data)*. Santa Clara, CA, pp. 821–826.

Landim, T., Alves-Souza, S.N. & DeSouza, L.S., 2013. Improving a website with web analytics — A case study. In *8th Iberian Conference on Information Systems and Technologies (CISTI)*. Lisbon, Portugal: IEEE, pp. 1–5.

Lin, J. & Ryaboy, D., 2013. Scaling big data mining infrastructure. *ACM SIGKDD Explorations Newsletter*, 14(2), p.6.

Linden, G., Smith, B. & York, J., 2003. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1), pp.76–80.

Long Jin et al., 2013. Understanding user behavior in online social networks: a survey. *IEEE Communications Magazine*, 51(9), pp.144–150.

Ma, H. et al., 2011. Recommender systems with social regularization. In *Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11*. New York, New York, USA: ACM Press, p. 287.

Madden, S. /Massachusetts I. of T., 2012. to Big Data. *IEEE Internet Computing*, pp.4–6.

Maniktala, M. et al., 2015. Finding the most informational friends in a Social Network based Recommender System. In *Annual IEEE India Conference (INDICON)*. New Delhi, pp. 1–6.

Oliveira, A.D.R. et al., 2012. Trust-based recommendation for the social Web. *IEEE Latin America Transactions*, 10(2), pp.1661–1666.

Prando, A.V. & Alves de Souza, S.N., 2016. Modular Architecture for Recommender Systems Applied in a Brazilian e-Commerce. *Journal of Software*, 11(9), pp.912–923.

Quirino, G.Z., Mals, N.P., Groterhorst, V.M., de Souza, S.N.A., et al., 2015. Meneduca — Social school network to support the educational environment. In *2015 Latin American Computing Conference (CLEI)*. IEEE, pp. 1–6.

Rennie, J.D.M. et al., 2003. Tackling the Poor Assumptions of Naive Bayes Text Classifiers. *IN PROCEEDINGS OF THE TWENTIETH INTERNATIONAL CONFERENCE ON MACHINE LEARNING*, pp.616--623.

Ricci, F., Rokach, L. & Shapira, B., 2011. Introduction to Recommender Systems Handbook. In *Recommender Systems Handbook*. Boston, MA: Springer US, pp. 1–35.

Rios, L.G. & Diguez, J.A.I., 2014. Big Data Infrastructure for analyzing data generated by Wireless Sensor Networks. In *2014 IEEE International Congress on Big Data*. IEEE, pp. 816–823.

Rokach, L. & Maimon, O., 2007. *Data Mining with Decision Trees*, WORLD SCIENTIFIC.

Saif, H. et al., 2016. Contextual semantics for sentiment analysis of Twitter. *Information Processing & Management*, 52(1), pp.5–19..

Schein, A.I. et al., 2002. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '02*. New York, New York, USA: ACM Press, p. 253.

Shani, G. & Gunawardana, A., 2011. Evaluating Recommendation Systems. In *Recommender Systems Handbook*. Boston, MA: Springer US, pp. 257–297.

Singh, S. & Singh, N., 2012. Big Data analytics. In *2012 International Conference on Communication, Information & Computing Technology (ICCICT)*. IEEE, pp. 1–4.

Sun, J. et al., 2015. Mining affective text to improve social media item recommendation. *Information Processing & Management*, 51(4), pp.444–457.

Xia, R. et al., 2016. Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis. *Information Processing & Management*, 52(1), pp.36–45.

Yang, S.-H. et al., 2011. Collaborative competitive filtering. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR '11*. New York, New York, USA: ACM Press, p. 295.

Zhang, Y. & Pennacchiotti, M., 2013a. Predicting purchase behaviors from social media. In *Proceedings of the 22nd international conference on World Wide Web - WWW '13*. New York, New York, USA: ACM Press, pp. 1521–1532.

Zhang, Y. & Pennacchiotti, M., 2013b. Recommending branded products from social media. In *Proceedings of the 7th ACM conference on Recommender systems - RecSys '13*. New York, New York, USA: ACM Press, pp. 77–84.