Cold-start Problem of Mobile News Client with Personalization Recommendation

Jun Li[1, a], Zhixin Shi[2, b], Jingang Liu[1, c] and Gao Lu[3, d]

[1]College of Information Engineering, Capital Normal University, Beijing 100048, China

[2]Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China

[3]The third academy of china aerospace science and industry corporation, Beijing 100074, China

[a]599298542@qq.com, [b]shizhixin@iie.ac.cn, [c]liujg2000@163.com, [d]ikelogo@gmail.com

Keywords: Personalization recommendation; Cold-start; User feature; Start-up system

Abstract. With news clients appeared in different fields, start-up systems, due to lack of user data, will cause the cold-start problem of personalization recommendation service. In the cold-start problem, this paper proposes a data migration-based cold-start algorithm, a cold-start algorithm based on user mobile device interest preference and a location services-based cold-start algorithm. Through analysis of performance differences and defects existed of these algorithms, so as to provide a reference of selection algorithms to users in solving the cold-start problem.

Introduction

News as a window of learning about the world for people and a leisure entertainment media, along with the rapid development of Mobile Phone Network (Such as 3G, 4G, etc.) and Mobile Internet [1], the news model has undergone tremendous changes, especially more and more different types of news clients in different fields appeared. Start-up systems [2], due to lack of user information and user history behavior data, can't provide personalized services to users. Therefore, how to deal with the cold-start problem caused by a small amount of data becomes one of difficulties in personalized recommended.

Cold-start Introduction

In a study show that if a user enters system provide valid personalized service, user loyalty and satisfaction will significantly enhance [1], therefore, cold-start problem of personalized recommendation that should be solved quickly. Among them, cold-start problem of recommendation system is one of difficulties, which divided into three aspects:

1) User cold-start;
2) Item cold-start;
3) System cold-start.

User cold-start or item cold-start of recommendation system is symmetrical problem, which can use similar methods to solve. Domestic and foreign studies have focused on item cold-start or user cold-start [3, 7], however, system cold-start actually is rarely considered. Therefore, system cold-start is more challenging than user cold-start or item cold-start. This paper proposes based on cold-start algorithm of data migration, based on cold-start algorithm of user mobile device interest preferences, location-based service cold-start algorithm which solve cold-start problem, subsequent sections will be discussed in detail.

Based on Cold-start Algorithm of Data Migration

Currently, Tencent, Sina and other large social networking sites already have hundreds millions of users, and provide data through calling interface for all kinds of system. Among them, start-up system through user authorization obtains social networking sites data so as to solve start-up system due to lack of data fail to provide user personalized services problem.

When user logs in start-up system, the system obtain user basic data, crawelling recent participation interactive of user (Such as users forwarding information, information reviews of friends and publish information's thumb and so on)**.** System analysis interactive information, which obtains user preferences and extracting user features, so as to accomplish users personalized recommendation service**.**

The algorithm mainly elaborate user feature, news feature, similarity calculation and so on**.**

User Feature. User feature extraction includes user fundamental features and user interaction characteristic**:**

1) user fundamental feature

User fundamental data includes nickname, sex, age, place of birth, date of birth, etc., user fundamental data features formula (2.1.1):

$$\text{fundament} = \text{sex} \times h + \text{pbirth} \times q \tag{2.1.1}$$

In this formula: fundament : user fundamental data features, sex: user's sex, h : sex's weight, pbirth: user's place of birth, q : pbirth's weight.

2) user interaction characteristic

Using natural language processing analyse interaction information to extract user interaction information characteristic. User interaction characteristic extracts interaction information that mainly is keywords and weights, User interaction characteristic's formula (2.1.2):

$$mulInfo = \sum_{i=1}^{m} keyword \times keyweight, m > 1 \tag{2.1.2}$$

In this formula: $mulInfo$: user interaction characteristic, m: concluding word number, $keyword$: key word, $keyweight$: the weight of key word.

User fundamental features and user interactive characteristic using formula to combine into user features. Formula (2.1.3):

$$userChar = \text{fumdament} \times \mu + mulInfo \times v \tag{2.1.3}$$

In this formula: $userChar$: user features, u : user fundamental data features's weight, v : user interaction characteristic's weight.

News Feature. Start-up system contains only basic information, which can use the Chinese word extract news feature.

Using the Chinese word extract article keywords so as to use TF-IDF [8] to calculate the weight of keywords, formula (2.2.1)、(2.2.2)、(2.2.3). The result of keywords generate word vector by weight in descending order, to obtain news of characteristics of each news.

$$tf(t,d) = \frac{f_{t,d}}{\max\{ f_{t',d} : t' \in d \}} \tag{2.2.1}$$

In this formula: $tf(t,d)$ is term frequency, $\max\{ f_{t',d} : t' \in d \}$ is maximum raw frequency of any term in the document.

$$idf(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|} \tag{2.2.2}$$

$idf(t,D)$: inverse document frequency, N : total number of documents in the corpus, $|\{d \in D : t \in d\}|$ number of documents.

$$tfidf(t,d,D) = tf(t,d) \bullet idf(t,D) \tag{2.2.3}$$

$tfidf(t,d,D)$: term frequency–inverse document frequency.

Similarity Calculation. Section 2.1 and section 2.2 can get news feature and user characteristics, if one can use cosine similarity formula calculate news feature and user characteristics [10], cosine similarity formula (2.3.1):

$$w_{u,j} = \frac{d_u \bullet d_j}{\sqrt{|d_u| \bullet |d_j|}} \qquad (2.3.1)$$

$w_{u,j}$: Result of similarity calculation, d_u: user u's feature, d_j: news j's characteristics.

After getting the result of similarity calculation, according to similarity score descending recommended for users of information in order to reduce the impact of cold-start of personalized recommendation service.

Algorithm Defect of Data Migration. Using data on migration log in all kinds of network system to people's daily life is made easier for different site system, user can be convenient login, can quickly access the underlying data can also be synchronized with user's dynamic information Social Network Site platform. There are corresponding defects, easily lead users to become "an account mountebank", to which there is a big risk. If Weibo account was closed, then the corresponding account cannot log in, user have to re-register and the previous information data is lost. Even worse is theft of user's funds.

Cold-start Algorithm of User Mobile Device Interest Preferences

Start-up system can get name of user's installing APP through authorization of user, extracting APP features, furthermore, all the APP features constitute of user feature. Then can use the similarity formula to calculate user's similarity, which parameters are user features and news features, the result brings users personalized recommendation list.

This algorithm mainly discusses user features, news features, similarity calculation and defect of algorithm, which extract news features are similar to section 2.2, and is not described.

User Features. There are two ways of achieving APP characteristic, One is to get the name of installing APP, crawling APP profile and using semantic analysis to achieve APP features; Another way gets popular APP by application store, crawling APP profile so as to semantic analysis gets APP characterized, when users use start-up system which gets APP name and directly generates APP feature. Whichever way use all of APP feature should be together to generate user characteristics. For example, if a user has installed Beauty Camera, Aunt, crawler crawl Beauty Camera profile and Aunt profile, so as to obtain Beauty Camera feature and Aunt feature, and then Aunt features and Beauty Camera feature integration, ultimately generate user characteristics.

User feature extraction process includes obtaining user installing APP, crawling APP Profile, generating APP features, integration user features four steps. Among them, crawling APP profile and generating APP features will occupy more than 90% of time, it will get crawling APP profile and generating APP features a two-step process in advance, and it will shorten time to get user characteristics, in order to improve customer satisfaction and loyalty degree.

Similarity Calculation. Start-up system analyses news, and later extract news features, this algorithm extract news characteristic is as same as 2.2 sections.

With above way obtained user's preference and user characteristic so as to calculate user interest. seeing similarity calculation formula (2.2.1).

After getting user's similarity scores, it can generate a similarity score descending recommended list, when user logs in the system will list recommended to user, thereby reducing impact of cold-start of personalized service.

Algorithm Defect of User Mobile Device Interest Preferences. Currently, Apple's App Store's applications has more than 1.5 million, while the number of applications Android platform more than a large number of Apple's App Store, So there is a great degree of difficulty which is analysed in APP features.

In addition, start-up system access to mobile phone which may involve user privacy problem, if

users don't want to authorize and news system cannot provide personalized recommendations to users to use the phone in the interest preferences.

Location-based Service Cold-start Algorithm

The Pew study found that 58% of smart phone users using location-based service, which are used in navigation and location-recommended 55% [11]. Location-based service [12] are through with mobile terminals, wireless and satellite communication network and determines mobile user's location to provide location service user. Liu et al found that [13]: In different regions, presence of news trends of interest to its geographical location and user's related to some of the differences. Location-based information referral service is a GPS data-driven social networking service, and people can share history of location information to share life experiences.

So, get the user's location to solve the problem of cold-start of great help, but also can improve system user satisfaction and loyalty.

User Features. User can accurately obtain current location (eg. West Third Ring Road, Haidian District, Beijing) when use smart phone. When get geographical location, it can be divided into provinces (autonomous regions , municipalities), city (prefecture) and county (banner, district), and other four levels, but each level can be given appropriate weight, table 1:

Table 1   Geographical level allocation table

	weight(scores)
provinces(autonomous regions , municipalities)	1
city (prefecture)	2
county (banner, district)	3
other	4

System gets the location information, analyzing location information obtained user characteristics. Such as West Third Ring Road, Haidian District, Beijing, It can be identified as user feature vector word {Beijing: 1, Haidian: 3, West Third Ring Road: 4}.

News Feature. Start-up system uses the Chinese word extraction news place-name by using TF-IDF calculate over name of keyword weight and get news features. After extracting place-name, place-name is compared with national place-name standard library to obtain news feature. When extracting place-name, it should pay attention to the same meaning of expression but not the same words, such as the capital and Beijing, RamCity and Guangzhou, Chongqing and Mountain City, which are described the same place.

Similarity Calculation. With above way obtains user preference and user characteristic so as to calculate user interest, seeing similarity calculation formula (2.2.1). After getting user's similarity scores, it can generate a similarity score descending recommended list, when user logs in the system will list recommended to user, thereby reducing impact of cold-start of personalized service.

Algorithm Defect of Location-based Service. This algorithm relates to user's privacy problem, when the user doesn't want to provide their location, and system can't provide personalized services.

In some places, news is little which can only recommend other place news. This would allow user satisfaction with system drops, causing users to no longer use the system.

Conclusion

In order to improve the cold-start problem of personalization recommendation, this paper proposes 3 algorithms: a cold-start algorithm based on data migration, a cold-start algorithm based on user mobile device interest preferences and a location service-based cold-start algorithm. In these algorithms, using Chinese word segmentation and semantic analysis to extract users features and news characteristics. Using similarity formula to calculate the user features and news characteristics, to get a recommendation list of user features, so as to reduce the system cold-start impact on customer satisfaction and loyalty. In the solving of the cold-start problem, it not only needs to

0976

innovate and improve on methods, but also pays attention to users' participations and feedbacks.

References

[1] Li Cong. Research on collaborative filtering bottleeneck in e-commerce recommendation system [D]. Hefei: Doctoralthesis. Hefei University of Technology, 2009.

[2] Schein A I, Popescul A, Ungar L H,et al.Methods and metrics for cold-start recommendations [C]//Proceedings of the 25th annual international ACM SIGIR conference on research and devlopment in information retrieval.ACM,2002:253-260

[3] Park Y J, Tuzhilin A. The long tail of recommender systems and how to leverage it[C]. In Proceedings of the 2008 ACM Conference on Recommender Systems, ACM, 2008: 11-18.

[4] Martinez L, Perez L G, Barranco M J. Incomplete preference relations to smooth out the cold-start in collaborative recommender systems [C].In Fuzzy Information Processing Society, NAFIPS 2009,Annual Meeting of the North American,IEEE,2009: 1-6.

[5] Chen Ke-han, Han Pan-pan, Wu Jian. User clustering based social network recommendation [J].Chinese Journal of Computers, 2013, 36(2) : 349-359.

[6] Wang Peng, Wang Jing-jing, Yu Neng-hai. A kernel and user-based collaborative filtering recommendation algorithm [J].Journal of Computer Research and Development, 2013, 50(7) : 1444-1451.

[7] Zhu Rui,Wang Huai-min, Feng Da-wei. Trust worthy services selection based on preference recommendation [J].Journal of Software, 2011, 22(5): 852-864.

[8] Gerard Salton, Christopher Buckley. Term 一 weighting Approaches in Automatic Text Retrieval[J] . Information Processing & Management,1988, 24(5): 513-523.HC Wu, RPW Luk, KF Wong, et al. Interpreting TF-IDF term weights as making relevance decisions [J]. Acm Transactions on Information Systems, 2008, 26(3):55-59

[9] HC Wu, RPW Luk, KF Wong, et al. Interpreting TF-IDF term weights as making relevance decisions [J]. Acm Transactions on Information Systems, 2008, 26(3):55-59

[10] Pew Research. 28% of U.S. adults use mobile and social location-based services. 2011. http://pew esearch.org/pubs/2096/mobilesocial-location-based-services-geosocial-social-media-location-tagging

[11] Jensen CS, Christensen AF, Pedersen TB. Location-Based services: A database perspective. In: Proc. of the Scandinavian Research Conf. on Geographical Information Science. New York: ACM Press, 2001. 59-68.

[12] Liu J,Dolan P.Personalized news recommendation based on click behavior[C].Proceedings of the 15th International Conference on Intelligent User Interfaces,ACM,2010: 31- 40.

[13] GEDIK B, LIU Ling. Location privacy in mobile systems: a personalized anonymization model[C]/ /Proc of the 25th IEEE International Conference on Distributed Computing Systems.2005: 620-629.