TAXONOMY-DRIVEN CONTENT-BASED RECOMMENDATION FOR NEW ITEMS

THALES FILIZOLA COSTA

TAXONOMY-DRIVEN CONTENT-BASED RECOMMENDATION FOR NEW ITEMS

Dissertation presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Nivio Ziviani
Co-Advisor: Rodrygo Luis Teodoro Santos

Belo Horizonte
February 2014

Costa, Thales Filizola

C837t Taxonomy-driven content-based recommendation for new items / Thales Filizola Costa. — Belo Horizonte, 2014.
xiv, 47 f.: il.; 29cm.

Dissertação (mestrado) - Universidade Federal de Minas Gerais – Departamento de Ciência da Computação.

Orientador: Nivio Ziviani
Coorientador: Rodrygo Luis Teodoro Santos

1. Computaçâo - Teses. 2. Recuperaçâo da informaçâo - Teses. 3. Sistemas de recomendaçâo - Teses.
I. Orientador. II. Título.

519.6*73(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Taxonomy-driven content-based recommendation for new items

THALES FILIZOLA COSTA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

Nivio Ziviani

Prof. Nivio Ziviani - Orientador
Departamento de Ciência da Computação - UFMG

RLSantos

Prof. Rodrygo Luis Teodoro Santos - Coorientador
Departamento de Ciência da Computação - UFMG

Adriano Vel

Prof. Adriano Alonso Veloso
Departamento de Ciência da Computação - UFMG

WMeira

Prof. Wagner Meira Júnior
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 21 de fevereiro de 2014.

Acknowledgments

I would like to begin this acknowledgments by thanking god for his blessings. Next, I am thankful for the people who have supported me throughout my entire life, from primary, through high school, college and eventually my masters degree: my parents, Wanduir and Dirce, and my sister, Thabata. Without you none of this would be possible. I appreciate all the advice and late night snacks. Thank you. Also, I am also thankful to all my other relatives, uncles, aunts and cousins. You were important despite not being with me on a day-to-day basis.

In addition, I am very grateful for the support of my girlfriend, Thalia. Thank you love, for all your patience and help along the long nights reading and discussing this thesis. Without you I would not have dispended so much enthusiasm and energy on this thesis. Thanks, love you.

Also, I would like to say thanks for my advisor Nivio, who taught me you cannot achieve anything without hard work and excellence. Thanks to my co-advisor Rodrygo, who helped me a lot in the process of conducting this research and writing this thesis. For both of you, I am pleased for all the shared knowledge. Moreover, I am thankful for the other member of my dissertation board, Wagner and Adriano for their valuable advices.

Finally, I am very thankful for all the friends I made during my journey. My classmates and friends: Eduardo, Luis, Mateus, Assahaf, Caio, Artur, Mattar and others, who stayed at my side during the four (long) years of graduation. I am also thankful for my co-lab: Anisio, Itamar, Cristiano, Aline, Aécio, Leonardo, Adolfo, Wallace, Sabir, Alan, Rickson, Wladmir and others that eased the work, making the hours in laboratories more funnier. Last but not least, I want to say thank you for my colleagues from CAMPS: Geraldo, Raphael, Pesce, Las Casas, Vinicius, Gabriel, Evandro and others for their chat time and fellowship. To conclude, I want to say thank you for others who contributed in some way. I am sorry that I have not quoted yours names, but thanks anyway.

"Wisdom is better than foolishness in the same way that light is better than darkness.
Wise people use their minds like eyes to see where they are going.
But for fools, it is as if they are walking in the dark."
(Ecclesiastes 2:13-14 (ERV))

Abstract

Recommender systems aim at predicting the preference of a user towards a given item (e.g., a movie, a song, a news story). Effective recommendations can be produced through collaborative filtering, in which case the previously manifested preferences of a community of users are leveraged to inform the recommender system. For systems that must cope with continuously evolving item catalogs, there will be a considerable rate of new items for which no past preference is known. This problem, known as the new item recommendation problem, may hamper the performance of recommender systems that are based solely on collaborative filtering. To overcome this problem, we propose an information-theoretic approach that exploits content-based features derived from taxonomies associated with the cataloged items. In contrast to previous content-based approaches, our approach is domain-agnostic, and can be directly deployed to produce effective recommendations for new items in different domains. For domains where an explicit taxonomy is not available, we show that a suitable one can be derived implicitly using Latent Dirichlet Allocation. Our experiments using two publicly available datasets attest the effectiveness of the proposed approach, which significantly outperforms state-of-the-art baselines from the literature.

Keywords: Recommender Systems, Cold-start Problem, Taxonomy, New User, New Item.

Contents

Acknowledgments **vii**

Abstract **xi**

1 Introduction **1**

1.1 Thesis Statement 2

1.2 Dissertation Contributions 2

1.3 Dissertation Outline 4

2 Background **5**

2.1 Basic Concepts 5

2.1.1 Recommender Systems 5

2.1.2 The Cold-Start Problem 7

2.1.3 Taxonomies 8

2.2 Related Work 10

2.2.1 Cold-Start Item 10

2.2.2 Cold-Start User 11

2.2.3 Cold-Start System 12

2.2.4 Taxonomies in Recommendation 12

2.3 Summary 13

3 Proposed Method **15**

3.1 The NIT-BR Model 15

3.2 Item and User Representation 16

3.3 Term-Category Weighting 18

3.4 Summary 19

4 Experimental Setup **21**

4.1 Research Questions 21

4.2 Recommendation Datasets 22
4.3 Recommendation Baselines 23
4.4 Training and Evaluation Procedures 24
4.5 Parameters Setting 26
4.5.1 NIT-BR model 26
4.5.2 Baselines 28
4.6 Summary 29

5 Experimental Evaluation 31
5.1 Taxonomy Usefulness (Q1) 31
5.2 NIT-BR Effectiveness (Q2) 34
5.3 Automatically Generated Categories (Q3) 37
5.4 Summary 39

6 Conclusions and Future Work 41

Bibliography 43

Chapter 1

Introduction

Recommender systems are a widely adopted class of information systems designed to recommend items, such as products, movies and news stories, for a user or a group of users. These systems are a key component in today's large e-commerce companies, automating the process of personalization at a large scale, which enables individual personalization for each customer. Jeff Bezos, CEO of Amazon.com, once emphasized that the success of his company relies on providing a unique store for each user, as modern e-commerce companies must search for multiple products that meet the multiple needs of multiple costumers [Schafer et al., 1999]. This constant need for satisfying each user, associated with a strong competition, demands an increasing number of products in offer. Therefore, recommender systems have to deal with the continuous evolution of their item catalog as well as their user base, which leads to the problem of how to provide effective recommendations for new items and/or new users.

This so-called *cold-start problem* [Anand and Griffiths, 2011; Park and Chu, 2009] may hamper the performance of recommender systems, as these systems are unable to draw any inference of the preferences of a given user for a particular item. In this dissertation, we focus on the cold-start problem for items—i.e., the lack of ratings for an item—also known as the *new item recommendation problem* [Adomavicius and Tuzhilin, 2005; Cremonesi et al., 2011]. Formally, a new item i is provided to the recommendation method as input, knowing only its content-like features, such as description and associated taxonomy categories. Since this item is new, none feedback (ratings) from users are provided. The method subsequently outputs the top ranked users R_u most likely to be attracted towards item i. Effectively tackling this problem is critical for the success of recommender systems as the owners of such systems have an interest in reducing the *item latency*, which is the time between the release of a new item and its first appearance within a recommendation list [Anand and Griffiths, 2011]. At the same time, as with any other item, in order to sustain *customer loyalty*, the recommendation of new items must also be surprising and effective [Sarwar et al., 2000; Schafer et al., 1999].

Collaborative filtering (CF) algorithms are generally reported to have the best accuracy in traditional recommendation scenarios [Cremonesi et al., 2010; Deshpande and Karypis, 2004]. However, these algorithms cannot cope effectively with the new item recommendation problem [Adomavicius and Tuzhilin, 2005; Schein et al., 2002], when there is not enough rating information to model the users' preferences towards new items. With the lack of ratings, an alternative approach to tackle the new item recommendation problem is to exploit features based on the contents of the new item. In particular, existing content-based (CB) recommendation approaches (e.g., [Furnas et al., 1988; Gunawardana and Meek, 2008; Pilászy and Tikk, 2009]) typically leverage domain-specific features, such as cast and director for movies, or author and publisher for books. On the other hand, these approaches have a clear limitation when generalizing to different domains, as different domains have different attributes of interest.

In order to overcome the limitations presented before, we propose a simple yet effective taxonomy-driven content-based approach. In particular, we model the new item recommendation problem as a traditional search problem, by relying on a term-based representation of items and users. Furthermore, we perform an information-theoretic term selection, which refines both representations by selecting semantically important terms from the description of items—e.g., to describe the movie "Titanic", the term "love" is more important then the term "collides". This selection is especially important in cold-start scenarios, where the lack of data increases the negative effect of noise on the quality of recommendations. In contrast to previous CF and attribute-based CB approaches, our approach requires no explicit users' preference and is easily ported to produce recommendations for different domains, including multimedia domains such as songs and videos.

1.1 Thesis Statement

The statement of this thesis is that the use of taxonomies to improve the representation of items and users is an effective approach for recommending new items to users.

1.2 Dissertation Contributions

The key contributions of this dissertation can be summarized as follows:

i) We introduce a supervised approach for recommending new items to users.

We propose a novel supervised approach, called *New Item Taxonomy-Based Recommender* (NIT-BR), which addresses the new item recommendation problem by relying on content-based features available on any domain, namely, the description and the taxonomy of items. In particular, our approach tackles this problem as a traditional search problem, by representing the descriptive terms of a new item as a query, and each user as a virtual document, comprising the terms of items that the user positively rated in the past.

ii) We exploit taxonomies to provide effective recommendations for new items.

We study the problem of recommending new items, for which no previous ratings are known. In particular, we use taxonomy as a discriminative source of evidence by selecting taxonomy-like terms from the description of items, which seeks to ensure that only terms that carry important information about a given item are selected, drastically reducing the feature space, limiting the noise and increasing the effectiveness of our model. In addition, we give insights on understanding why particular terms from the descriptions of items are more important than others. Furthermore, to extend the usage of our model for scenarios where an explicit taxonomy is not available, we propose an alternative approach that relies on topics generated by Latent Dirichlet Allocation (LDA) [Blei et al., 2003] as a replacement for missing taxonomy categories.

iii) We thoroughly evaluate the proposed approach using two publicly available datasets.

Our experimental results attest the effectiveness of our approach across two publicly available datasets for movie and book recommendations. In particular, the experiments conducted attest the effectiveness of NIT-BR compared to state-of-the-art baselines from the literature, even when no taxonomy is available. Moreover, we show that information-theoretic metrics, such as Mutual Information, are particularly suitable for selecting effective item descriptors. In addition, we meticulously investigate the robustness of our approach when applied in domains where an explicit taxonomy is not available, showing that while categories automatically derived using LDA can be used effectively by our model, the availability of an explicit, manually curated taxonomy can provide further gains.

1.3 Dissertation Outline

The remainder of this dissertation is organized as follows:

- Chapter 2 presents background information which will guide the understanding of the subsequent chapters. In particular, it begins by describing three key concepts—*recommender systems*, *cold-start problem* and *taxonomies*—and their terminology. Besides these concepts, this chapter also discusses the importance of the cold-start problem for recommender systems. Finally, this chapter reviews the literature related to the cold-start problem and to the use of taxonomies in recommendation.

- Chapter 3 provides an in-depth description of the NIT-BR model, our proposed approach to tackle the new item recommendation problem. In particular, this chapter first introduces our model. Next, it details our representation of items and users. Lastly, it describes the information-theoretic metrics that are used to score the importance of terms in a given taxonomy category.

- Chapter 4 details the experimental setup and presents the research questions addressed in this dissertation. In particular, we discuss the datasets, the baselines and the evaluation methodology used in our investigations.

- Chapter 5 discusses our experimental results, therefore answering our stated research questions. In particular, this chapter first investigates the usefulness of taxonomies within our model, hence addressing our first research question. Besides this investigation, we also present some insights on why taxonomy-related terms are more relevant to describe an item than others. Next, we focus on our second research question, by assessing the effectiveness of our model against state-of-the-art new item recommendation approaches from the literature. Lastly, this chapter evaluates the effectiveness of our model when an explicit taxonomy is not available, therefore answering our third research question.

- Chapter 6 provides a summary of the contributions and conclusions made throughout this dissertation. Finally, it presents directions for future research.

Chapter 2

Background

This chapter introduces background information to help understand our proposed solution for the new item recommendation problem and its impact in the literature. For convenience, we divide this chapter in two sections. Section 2.1 presents basic terms and concepts used along this dissertation. Some of these concepts were already presented in Chapter 1 and are now further detailed. Section 2.2 introduces previous literature that is, in some points, related to the problem, to the proposed solution, or to both. In addition, this second section helps to measure our impact in the literature by positioning our proposal against these previous works.

2.1 Basic Concepts

This section presents three basic concepts that will guide the understanding of the following chapters. With these concepts, we present some terminology used throughout this dissertation. In particular, Section 2.1.1 introduces the concept of *recommender systems*. Section 2.1.2 presents the *cold-start problem* and some classification from the literature about this problem. Finally, Section 2.1.3 defines and provides examples of *taxonomies*.

2.1.1 Recommender Systems

Recommender systems are a widely adopted class of information systems designed to recommend items (e.g., products, movies, news) for a user or a group of users [Resnick and Varian, 1997; Mahmood and Ricci, 2009; Burke, 2002]. In particular, these systems are classified as [Adomavicius and Tuzhilin, 2005]:

i) *Collaborative filtering* methods are based on past user ratings. To provide recommendations for a user, these methods first identify other users with similar taste to recommend items that those similar users have liked.

ii) *Content-based* methods exploit the content attributes of items, i.e. these methods recommend items with similar content to those previously liked by a given user.

iii) *Hybrid methods* are combinations of the two previous approaches.

Large online companies—such as Facebook, Google and Netflix—have shown that the development of an effective recommender system is important for their success, as these systems are now key components of their business models. In particular, Ricci et al. [2011] presented the following benefits of effective recommender systems:

- *Increase the number of items sold,* as these systems help users to find their needs and wants.

- *Sell more diverse items,* as these systems focus on serendipity, by presenting items that might be hard to find.

- *Increase the user satisfaction,* as the user experience will be improved by using such systems.

- *Increase user fidelity,* as effective recommendations positively affects the users' perception of the system.

- *Better understanding what the user wants,* as the owner of such systems have access to a personalized profile of users and can directly influence the system's management.

Besides this business interest, recommender systems are a growing research field, with focus on many diverse areas, e.g. "explaining recommendations", "user trust", "group recommendation", among others. In fact, conferences targeted to these systems, such as the "ACM Recommender Systems Conference", have been recently created.

In this dissertation, we study the use of content-based methods on the cold-start problem scenario. More specifically, we investigate the use of taxonomies as a discriminative evidence for cold-start recommendation, by proposing a taxonomy-driven content-based model, which we test for this problem.

2.1.2 The Cold-Start Problem

As discussed in Section 2.1.1, recommender systems perform personalized recommendations of items to users or groups of users. In order to achieve personalization, these systems require user feedback.[1] The cold-start problem happens when there are users or items without preference, i.e., when the lack of information (feedback) for users or items hampers the process of recommendation [Lam et al., 2008; Schein et al., 2002].

Although the definition presented before generalizes the cold-start problem for both items and users, this problem has been classified in prior literature [Schein et al., 2002; Park and Chu, 2009] as follows:

- ◇ *Cold-start user problem* is the lack of ratings from a user that hampers the recommendation of personalized items for this user. It is also called the *new user problem.*

- ◇ *Cold-start item problem* is the lack of ratings from an item that reduces the effectiveness of its recommendation. It is also called the *new item problem.*

- ◇ *Cold-start system problem* is the lack of ratings from both users and items, being the most difficult scenario.

It is important to notice that almost any recommender system is likely to suffer from the three variants of the cold-start problem for the following reasons. First, every recommender system starts in some point where the amount of overall information (users' feedback) is small or even not available at all. A system in this context has difficulty in providing effective recommendations due to the lack of ratings from both users and items (cold-start system problem). One might suppose that once a system overcomes this initial stage, it does not suffer from cold-start anymore. However, an effective recommender is always seeking for new items to keep its catalog of items fresh and desirable, therefore maintaining users satisfied. By incorporating new items into the catalog, this system is susceptible to suffer from the new item problem (cold-start item). Finally, every successful recommender system has an increasing user base. With the attraction of new users comes the problem of providing effective recommendations for them (cold-start user).

Another classification exists for the cold-start problem. In particular, Park et al. [2006] and Park and Chu [2009] complemented, with a different focus, the classification presented before. This complementary classification is as follows:

[1] The user feedback can be either explicit (e.g., rating, review) or implicit (e.g., click through, search history, and purchase log).

- *Extreme cold-start* happens when users or items have no expressed preferences, being completely new to the system. This variation is important since purely collaborative filtering algorithms—which require previous information to build their model—are incapable of providing recommendations in this context.

- *Non-extreme cold-start* is the soft case. In this variation of the problem, items or users are considered new if they have a small amount of ratings or if they have appeared recently in the systems (e.g., a user who has been using the system for a few days, but has only a few ratings, is still considered new).

In this dissertation, we focus on the cold-start problem over items, also called the *new item recommendation problem*. This problem is of crucial importance for the success of recommender systems for the following reasons. First, the owners of such systems have an interest in reducing the *item latency*, which is the time between the release of a new item and its first appearance within a recommendation list [Anand and Griffiths, 2011]. The reduction of the item latency directly affects the revenue of these systems, as new items will be on the shelves. Second, the recommendation of a new item must also be remarkable and effective to increase *customer loyalty* [Sarwar et al., 2000; Schafer et al., 1999]. From the user point of view, a system capable of providing an effective recommendation of new items has the benefit of satisfying this user's needs with the most up-to-date items.

Furthermore, we are particularly interested in the extreme occurrence of this problem. As we will discuss in Section 2.2, the majority of previously published works developed domain-dependent solutions, which are suitable for one domain but have difficulty in generalizing for other domains. However, our intent in this dissertation is to craft a content-based solution that generalizes to different domains. To this end, we restrict our chosen features to only the description and taxonomy categories of items—features that are generally available in many domains.

2.1.3 Taxonomies

A taxonomy can be defined as a collection of categories according to which the items in the catalog of a recommender system can be organized [Cho and Kim, 2004; Weng et al., 2008]. Moreover, any item must belong to one or more taxonomy categories. For instance, in the taxonomy of Amazon,[2] presented in Figure 2.1, the book "City of Bones" belongs to the categories *Hard-Boiled* and *Mythology*.

[2] http://www.amazon.com

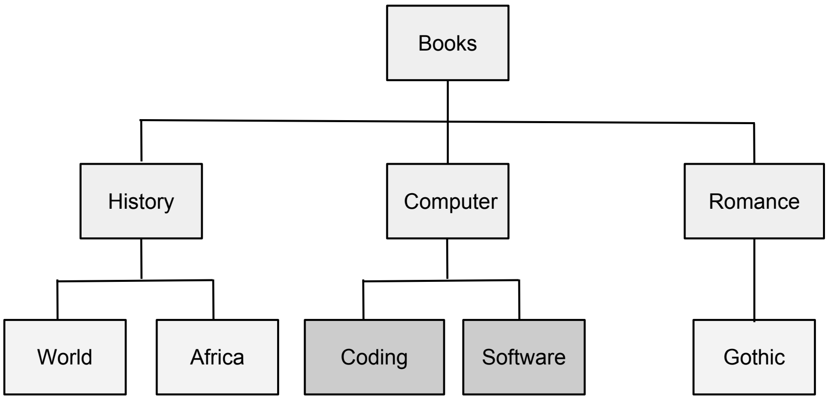


Figure 2.1. Illustration of a small part of Amazon's books taxonomy.

In some domains, the categories comprised by a taxonomy can be hierarchically organized in a tree-like structure, such as Amazon's taxonomy presented before. However, in this dissertation, we focus on flat taxonomies[3] and leave the exploration of hierarchical taxonomies for future work. In fact, this focus allows us to directly use a large available set of flat taxonomies, such as IMDB's[4] movie genre taxonomy presented in Figure 2.2.

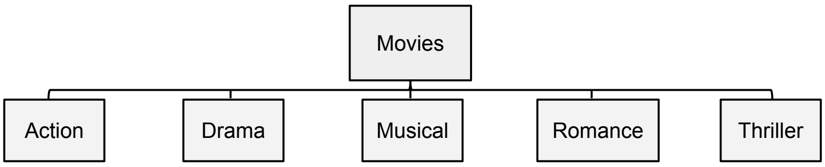


Figure 2.2. An excerpt of IMDB's movie genre taxonomy.

As we will show in the next section, taxonomies have been previously exploited in recommender systems in distinct ways. In addition, the literature gives some insights that help understand why this source of information is effective. According to some authors [Ahmed et al., 2013; Ziegler et al., 2004; Cho and Kim, 2004], taxonomies are assigned by humans and capture the knowledge about the domain, hence providing strong evidence about the semantics of items. Kanagal et al. [2012] also noticed that taxonomies can help in dealing with the cold-start problem because even though the catalog of items is highly dynamic, the taxonomy is relatively stable. Moreover, Weng

[3] When using hierarchical taxonomies, such as Amazon's book taxonomy, we exploit only the leaf categories, therefore resulting in a flat taxonomy.

[4] Internet Movie Database, http://www.imdb.com/genre/

et al. [2008] investigated the relation between users' preferences for items and their preference for a specific taxonomy category, concluding that users share not only similar item preferences but also similar taxonomic preferences.

In this dissertation, based on the definition of taxonomy and on the insights presented before, we exploit taxonomies as a complementary and potentially effective source of evidence, which embeds the semantics of items and may also help group users according to their preferences. More specifically, we exploit taxonomies as a domain-agnostic evidence to help identify the most informative terms in the description of each cataloged item, to improve the effectiveness of existing content-based approaches for recommending new items in different domains.

2.2 Related Work

Several approaches have been proposed in recent years to tackle the cold-start problem. In this section, we overview the most prominent of such approaches. In particular, we describe approaches dedicated to the three classifications of the problem (presented in Section 2.1.2), the cold-start problem for items, users, as well as a combination of both, referred to as the cold-start system problem [Anand and Griffiths, 2011; Park and Chu, 2009]. Finally, we cover approaches that also exploited taxonomies in some manner to improve recommendation effectiveness.

2.2.1 Cold-Start Item

The lack of ratings for new items may severely impact the effectiveness of a recommender system. To tackle the cold-start item problem, several approaches have been proposed in the literature. A traditional solution to this problem relies on the identification of users who have previously manifested an interest towards cataloged items with content similar to that of the new item. While such content-based approaches are generally effective in a cold-start scenario, they also have shortcomings [Lops et al., 2011]. In particular, word-level features may not capture the preferences of a user towards an item as well as explicit ratings would. In addition, domain-specific features (e.g., cast and director for movies, or author and publisher for books) may not generalize well across different domains. To overcome these limitations, alternative approaches have been proposed to exploit latent features of an item. The most prominent of such approaches—and the current state-of-the-art for the cold-start problem [Bambini et al., 2011]—is latent semantic analysis (LSA), which represents the cataloged items in a lower dimensional space of latent concepts. In contrast to these content-based approaches, we exploit features derived from the categories underlying any existing taxonomy of items under an information-theoretic recommendation model, as we will describe in Chapter 3. In our investigations in Chapter 5, LSA is used as a baseline.

As an alternative to content-based approaches, another family of approaches addresses a slightly relaxed version of the cold-start item problem, in which only a few ratings (as opposed to none) are available for a new item. In particular, this relaxation enables the use of traditional collaborative filtering (CF) algorithms. In this context, Cremonesi and Turrin [2009] showed that item-based CF [Sarwar et al., 2001] generally outperforms classical CF approaches based on singular value decomposition (SVD) [Sarwar et al., 2002]. Alternatively, several approaches have been proposed to exploit CF information to weigh content-based features: aspect models [Schein et al., 2002], Boltzmann machines [Gunawardana and Meek, 2008], association rules [Leung et al., 2008], and linear transformation [Pilászy and Tikk, 2009]. In addition, Adomavicius and Tuzhilin [2005] suggested the use of hybrid algorithms, which combine CF and content-based approaches, however, Park and Chu [2009] noted that even though hybrid algorithms are suitable for the cold-start problem, they focus on improving the overall accuracy of a recommender system rather than on improving their effectiveness for new items. In contrast to these approaches, we tackle the extreme version of the cold-start item problem, in which no ratings are available for the new item, a crucial scenario where CF approaches cannot be applied [Park et al., 2006].

Lastly, rating elicitation approaches have been proposed to select target users from the recommender system's user base who could provide ratings for a new item. For instance, Anand and Griffiths [2011] proposed a market-based approach to select the users with the highest influence spread in the user base. Likewise, Liu et al. [2011] used a matrix factorization approach to identify representative users and items. The approaches in this family differ from the aforementioned approaches mainly because of the interactive nature of the elicitation process. In our proposed approach, for instance, no explicit feedback is required from users.

2.2.2 Cold-Start User

New users can also pose a problem to a recommender system. Among the existing solutions for the cold-start user problem, Zhou et al. [2011] proposed to initially interview new users, i.e., first-time users were requested to rate a few items before they could receive recommendations. Similarly, Levi et al. [2012] also interviewed new users for hotel recommendations. However, instead of evaluating an initial set of hotels, the users were asked for contextual travel information (e.g., reason for the trip, budget, number of children), which was then used to identify similar trips by other users. As an alternative approach, Lam et al. [2008] presented an automatic solution that combined user information (age, gender, and job) with CF to provide recommendations for new users. In contrast, we exploit the semantics of taxonomies to refine the representation of items and users for the cold-start item problem, as we will discuss in Chapter 3.

2.2.3 Cold-Start System

A few works have also tackled the cold-start system problem, a scenario in which both items and users are new to the recommender system. In particular, Park and Chu [2009] proposed to exploit user and item attributes (e.g., user demographic information and item content) in a linear pairwise regression approach. Alternatively, Park et al. [2006] used automatic evaluation agents, called filterbots, to generate missing ratings and enable the use of CF algorithms. In contrast to these approaches, we make no assumptions regarding the availability of domain-specific item attributes or the availability of ratings for new items. Instead, we leverage taxonomies as a domain-agnostic evidence to recommend items that are completely new to the recommender system.

2.2.4 Taxonomies in Recommendation

Taxonomies have also been exploited in the literature to alleviate the cold-start problem in CF. For example, Cho and Kim [2004] used taxonomies to reduce the dimensionality of the ratings matrix to improve recommendations based on nearest neighbor searches. Ziegler et al. [2004] and Weng et al. [2008] represented users by their interest in categories from a taxonomy of products. Kanagal et al. [2012] proposed a taxonomy-aware latent factor model, which combines taxonomies and latent factors using additive models. Similarly, Ahmed et al. [2013] proposed a hierarchical additive model to exploit users' preferences towards attributes of the cataloged items across different categories. In contrast to these CF approaches, we adopt a pure content-based formulation to identify any descriptive terms (as opposed to, for instance, predefined attributes) that are informative of the taxonomy categories to which an item belongs. As a result, our approach can be directly applied to any recommendation domain, provided that a taxonomy is available or can be automatically inferred given the items' textual description.

2.3 Summary

In this chapter, we presented basic definitions and concepts that allow the reader to better understand the following chapters. First, we introduced recommender systems. Next, we defined the cold-start problem, describing its importance to recommender systems. Last, we introduced taxonomies, which are a key feature in our proposal. In addition, we discussed related work from the literature, linking these with the basic concepts presented before. In Chapter 3, we will present our approach to tackle the new item recommendation problem.

Chapter 3

Proposed Method

In this chapter, we present a taxonomy-driven new item recommender model, which we refer to as New Item Taxonomy-Based Recommender (NIT-BR). We first formally introduce our novel content-based recommendation model particularly suited for cold-start scenarios. Next, we describe our choices for representing items and users within our model. Lastly, we describe several information-theoretic metrics to select taxonomy-related terms for an improved item and user representation.

3.1 The NIT-BR Model

The new item recommendation problem, illustrated in Figure 3.1, can be stated as the problem of finding which users should be recommended an item that is completely new to a recommender system.

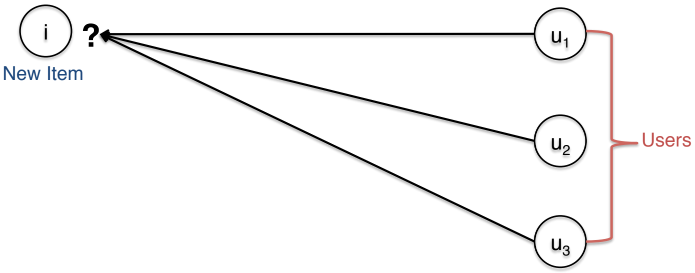


Figure 3.1. New item recommendation problem overview.

To tackle this problem, we introduce New Item Taxonomy-Based Recommender (NIT-BR), a novel content-based recommendation model that builds upon a classical information retrieval formulation. In particular, as Figure 3.2 shows, NIT-BR represents the new item as a "*query*", and each candidate user as a "*document*" that is potentially "*relevant*" for the new item.

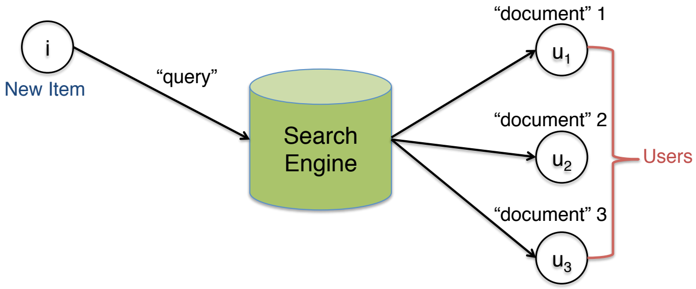


Figure 3.2. An overview of the NIT-BR model.

More precisely, given a new item i and a user u, NIT-BR estimates the relevance of u given i according to:

$$score(i,u) = \sum_{t \in \hat{\imath}} \Big(1 + \log(tf_{t,\hat{u}})\Big) \times \log\left(\frac{n}{n_t} + 1\right), \tag{3.1}$$

where $\hat{\imath}$ and $\hat{u}$ are term-based representations of the item i and the user u, respectively, $tf_{t,\hat{u}}$ is the frequency of the term t in $\hat{u}$, n_t is the number of users whose representation include t, and n is the total number of users in the system. Note that Equation (3.1) is a simple yet effective TF-IDF formulation operating on top of the term-based representations of items and users [Salton and Buckley, 1988]. While alternative formulations are certainly possible, we leave these for future investigations and instead explore multiple alternatives for effectively representing items and users, as discussed next.

3.2 Item and User Representation

In order to represent the items and the users of a recommender system in our NIT-BR model, we adopt a "domain-agnostic" term-based representation, which improves the generality of our model. In particular, as illustrated in Figure 3.3, a new item i is represented by a selection of the terms contained in its description. A user u, in turn, is represented by a selection of the terms contained in the description of the items that the user has positively rated in the past (i_1, i_2, and i_3 in the figure). The resulting representations of the new item and the user ($\hat{\imath}$ and $\hat{u}$, respectively) are then used as input to the scoring function defined in Equation (3.1).

To ensure we have meaningful representations for both the new item i and each user u, we propose to exploit the taxonomies associated with the cataloged items. In

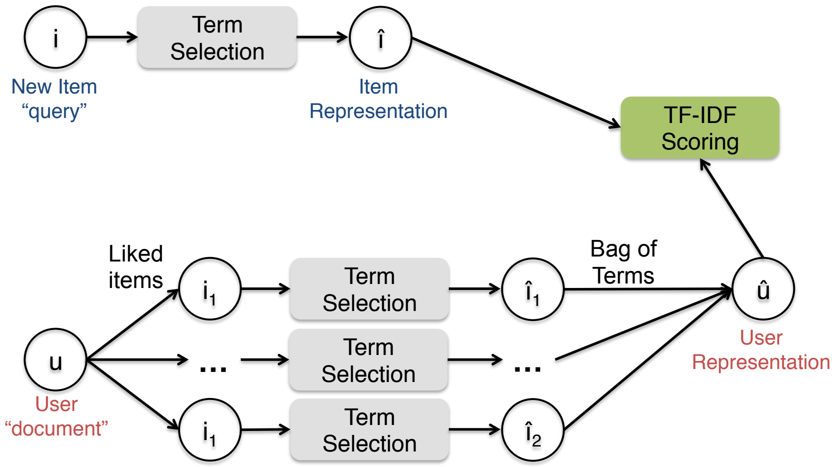


Figure 3.3. NIT-BR's top-down scheme.

particular, taxonomies provide a rich and domain-agnostic categorization of items into higher level concepts, which could better explain the interest of a user for a particular item. For instance, in the previous example, the interests of a user who purchases the book "City of Bones" are arguably better represented by terms related to the category *Mithology* than to ordinary terms such as "bones". The key challenge then becomes to distinguish taxonomy-related terms from ordinary terms given the description of each cataloged item.

More formally, given an item i whose description comprises a set of terms I, and a taxonomy C of categories defined over the entire item catalog, the term-based representation $\hat{\imath}$ of the item i can be defined according to:

$$\hat{\imath} = \underset{T \subseteq I}{\arg\max} \sum_{t \in T} \sum_{c \in C} w(t, c), \text{ s.t. } |T| \leq m, \tag{3.2}$$

where $T \subseteq I$ is a subset of the terms in I, and m is the maximum number of terms to be selected, which is a parameter of NIT-BR. Most notably in Equation (3.2), $w(t, c)$ defines the weight of the term t for each category $c \in C$. In Section 3.3, we describe several information-theoretic metrics that are used as alternative term-category weighting schemes in our experiments. Also note that, when $\hat{\imath} = I$, the representation of the item i includes all terms in the description of this item, which serves as a baseline in our investigations in Section 5.1 for validating the various alternative weighting schemes.

Lastly, in order to represent a user u, we concatenate the representation of all the items $i \in R_u^+$ that the user has positively rated in the past, according to:

$$\hat{u} = \bigcup_{i \in R_u^+} \hat{\imath}. \tag{3.3}$$

3.3 Term-Category Weighting

To weigh the informativeness of each candidate term t from the description of an item i with respect to each category $c \in C$ covered by the item, we investigate the effectiveness of four classical information-theoretic metrics. In particular, these metrics have previously been shown to be effective as term-weighting schemes in different scenarios, e.g., pseudo-relevance feedback [Brandão et al., 2013] and text classification [Liu et al., 2009]:

- **Pearson's CHI-squared (CHI2):** Measures the relationship between an expected frequency in the general population and an observed frequency. In our case, the expected frequency is the term likelihood $p(t)$, and the observed frequency is the probability of t given each category c, i.e., $p(t|c)$. It is given by:

$$w(t,c) = \frac{\big(p(t|c) - p(t)\big)^2}{p(t)}. \tag{3.4}$$

- **Dice's Coefficient (DICE):** Measures the similarity between two sets. We use it to evaluate the similarity between E_t and E_c, the sets of items related to the term t and the category c, respectively. If these sets are similar, t is considered closely related to c. This metric is defined as follows:

$$w(t,c) = 2 \times \frac{|E_{t,c}|}{|E_t| + |E_c|}. \tag{3.5}$$

- **Kullback-Liebler Divergence (KLD):** Also known as *relative entropy*, this metric estimates the difference between two probability distributions, $p(x)$ and $q(x)$. In our case, these distributions are $p(t|c)$ and $p(t)$, respectively. It is defined as:

$$w(t,c) = p(t|c) \times \log\left(\frac{p(t|c)}{p(t)}\right). \tag{3.6}$$

- **Mutual Information (MI):** Also known as *transinformation*, this metric quantifies the mutual dependence between two random variables X and Y, i.e., the information that X and Y share. In our case, X is E_c and Y is E_t, i.e., the greater the amount of shared information between X and Y, the more t is closely related to c. We calculate it as follows:

$$w(t,c) = \log\left(\frac{|E_{t,c}|}{|E_t| \times |E_c|}\right). \tag{3.7}$$

3.4 Summary

This chapter described NIT-BR, a novel content-based approach for the new item recommendation problem. NIT-BR models this problem as a classical information retrieval problem, by representing the new item as a "query" and each candidate user as a "document" that is potentially "relevant" for the new item. To further refine the representation of both items and users, we adopt an information theoretic approach and identify terms that are informative in light of the taxonomy categories associated with each item. By relying on generally available textual features, not only does NIT-BR effectively address the lack of ratings for new items, but it is also agnostic to any particular recommendation domain, as we will show in the experiments in the upcoming sections.

Chapter 4

Experimental Setup

In this chapter, we detail the experimental setup that supports our investigations in Chapter 5. In particular, Section 4.1 states the research questions we aim to answer. Section 4.2 introduces the recommendation datasets used in our experimentation, along with some standard pre-processing steps performed. Section 4.3 discusses our baselines, specially the Latent Semantic Analysis [Bambini et al., 2011], a state-of-the-art content-based algorithm. Section 4.4 presents our training and evaluation procedures, which are focused on the top-n recommendation task. Finally, Section 4.5 deliberates about the process of parameter tuning for NIT-BR as well as our baselines.

4.1 Research Questions

In this dissertation, we aim to answer the following research questions:

Q1. How useful are taxonomies for improving the representation of items and users?

Q2. How effective is our proposed NIT-BR model for recommending new items to users?

Q3. How does the lack of an explicit taxonomy impact the effectiveness of NIT-BR?

In particular, Q1 measures the effective of selecting taxonomy-liked terms. Meanwhile, Q2 focuses on comparing our proposal against other alternatives from the literature. Finally, Q3 compares the performance of our method using an alternative approach for missing taxonomy categories.

4.2 Recommendation Datasets

We report our experimental results on two publicly available datasets, namely, Book-Crossing[1] and MovieLens-1M.[2] We chose these datasets for three reasons. First, these datasets come from popular recommendation system domains—books and movies. Second, they have different information density, as presented by the "Sparsity" entry in Table 4.1. Third, the available taxonomy for each domain is different. In the book domain there is a detailed, hierarchical and informative taxonomy available (Figure 2.1), whereas in the movie domain we found nothing but a flat genre taxonomy (Figure 2.2).

Dataset	**Book-Crossing**	**MovieLens-1M**
# Items	$5712^\star$	3706
# Users	$3786^\star$	6040
# Ratings	$\approx 206\text{k}^\star$	$\approx$ 1M
# Categories	$855^\star$	$18^\star$
Rating Range	$0 \sim 10$	$1 \sim 5$
Sparsity	99.046%	95.531%

Table 4.1. Statistics of the augmented Book-Crossing and MovieLens-1M datasets. Sparsity is the percentage of empty cells in the user-item ratings matrix. The $\star$ symbol denotes statistics affected by the augmentation step described in Section 4.2.

Book-Crossing is a book recommendation dataset, with users' ratings for books, while MovieLens-1M is a movie recommendation dataset, with users' ratings for movies [Miller et al., 2003]. Due to Book-Crossing's extreme sparsity, following standard practice for this dataset [Gedikli and Jannach, 2010; Julià et al., 2009; Zhang, 2008; Ziegler et al., 2005], we discard items with less than five ratings, as well as users who have rated less than five items. In addition, since most of the approaches investigated in this dissertation rely on content-based features for recommendation, we complement the Book-Crossing dataset with the description and category of each book, obtained from Amazon.com, further discarding books with no associated description. Likewise, we complement MovieLens-1M with the synopsis and genre of each movie, obtained from the Internet Movie Database (IMDB).[3] The augmented datasets, summarized in Table 4.1, will be available in the authors' homepage for experimental reproducibility.

[1] www.informatik.uni-freiburg.de/~cziegler/BX

[2] www.grouplens.org/node/73

[3] www.imdb.com

4.3 Recommendation Baselines

We evaluate the effectiveness of our NIT-BR model in comparison to the following baselines:

- **Top Popular User (TPU):** Previous results suggest that non-personalized algorithms may outperform other recommendation approaches in extremely sparse scenarios [Park et al., 2006]. As a result, we include TPU as a baseline that scores users proportionally to the number of items they have rated in the past. Formally:

  $$score(i, u) = |R_u^+|, \tag{4.1}$$

  where R_u^+ is the set of items positively rated by the user u in the training set.

- **Segmented Top Popular User (STPU):** As a variant to TPU, STPU scores users proportionally to the number of items they have rated from the same categories as the new item. Formally:

  $$score(i, u) = \sum_{c \in C_i} |R_{u,c}^+|, \tag{4.2}$$

  where C_i is the set of categories to which the item i belongs and $R_{u,c}^+$ is the set of items that belong to category c and that were positively rated by user u.

- **Latent Semantic Analysis (LSA):** LSA is a state-of-the-art content-based algorithm, which presents an alternative scheme for reducing the dimensionality of the term-based representation of cataloged items [Bambini et al., 2011]. In particular, it projects both items as well as users into a lower dimensional space obtained via singular value decomposition [Bambini et al., 2011; Cremonesi et al., 2011]. Using LSA, the score of a user u for an item i is computed as:

  $$score(i, u) = \mathrm{sim}(\tilde{\imath}, \tilde{u}), \tag{4.3}$$

  where $\tilde{\imath}$ and $\tilde{u}$ are the vector representations of the item i and the user u in the resulting space of latent factors, and $\mathrm{sim}(\tilde{\imath}, \tilde{u})$ is the cosine similarity between $\tilde{\imath}$ and $\tilde{u}$. We implemented LSA using the S-Space Package [Jurgens and Stevens, 2010], and set the number of latent factors to 2,000 through cross-validation, as described in Section 4.4.

- **LSA "taxonomy" (LSA$_{tax}$):** In a recent work, Bambini et al. [2011] extended the original LSA algorithm to leverage features other than textual terms. Following their approach, in order to have a strong baseline that also exploits taxonomy categories, we augment the item-term matrix of the original LSA algorithm with a new item-category matrix M, such that:

$$M = \omega B, \tag{4.4}$$

  where B is a binary matrix, indicating the membership of each cataloged item to each taxonomy category, and ω is a weight assigned uniformly to all categories when leveraged as features by LSA. Through cross-validation, we set $\omega = 3$ in our experiments. This extended formulation is henceforth referred to as LSA$_{tax}$.

- **NIT-BR "all terms" (NIT-BR$_{all}$):** To assess the effectiveness of exploiting taxonomies for the new item recommendation problem, as an additional baseline, we consider a variant of our NIT-BR model that does not perform term selection. Instead, it represents an item using its entire description. Formally, this variant replaces the item representation in Equation (3.2) with:

$$\hat{\imath} = I, \tag{4.5}$$

  where I is the set of all terms in the description of item i. This baseline variant, referred to as NIT-BR$_{all}$, is used in our experiments as a representative of content-based approaches that compute the similarity between items and users using their textual representation. Although simple, such approaches are also effective [Lops et al., 2011].

4.4 Training and Evaluation Procedures

In order to evaluate our NIT-BR model for the new item recommendation problem, we assess its effectiveness at ranking a set of users according to these users' interest for a new item. The evaluation methodology we adopt is symmetric to the one proposed by Cremonesi et al. [2010]. While the original methodology tests the ability of a system to recommend items relevant to a given user, our adaptation tests the ability of the system to recommend relevant users for a given (new) item.[4] In particular, we first

[4] This adaptation emphasizes our focus on assessing the recommendation effectiveness for new items, as the original methodology cannot ensure that every item will be associated with at least one relevant user.

train each recommendation approach using all ratings in the training set. Then, for each new item i in the test set, we create one test instance for each "relevant" user u, i.e., a user who has rated i above or equal a certain threshold τ.[5] Besides the input item i and a relevant user u, a test instance also contains 1000 randomly sampled "non-relevant" users $\{\bar{u}_1, \cdots, \bar{u}_{1000}\}$, i.e., users who have not rated i. Given the item i and its associated sample of users $\{u, \bar{u}_1, \cdots, \bar{u}_{1000}\}$, the goal of a recommendation approach is to rank the relevant user u ahead of the non-relevant users $\{\bar{u}_1, \cdots, \bar{u}_{1000}\}$ in the sample.

$$\underbrace{\text{recall}(N) = \frac{\#\text{relevant retrieved}}{\#relevant}}_{\text{Traditional recall}} \xrightarrow{\text{One relevant}} \frac{\#hits}{\#tests}$$

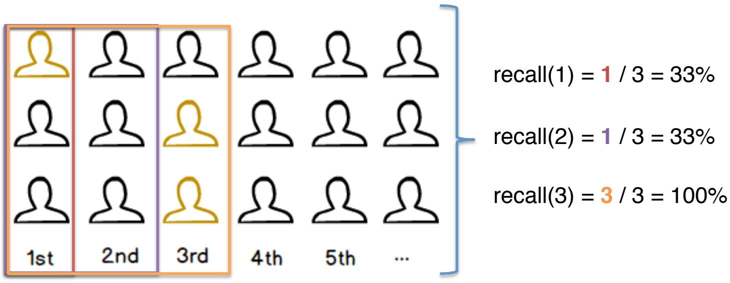


Figure 4.1. Formal definition of recall at N with one relevant, followed by an example.

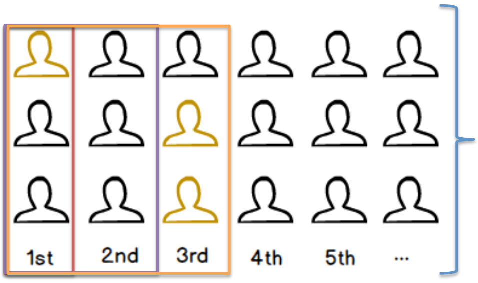
To assess the effectiveness of the various recommendation approaches considered in our investigation, we report recall and precision figures at different rank cutoffs [Baeza-Yates and Ribeiro-Neto, 2011]. In particular, for a given cutoff N, if the relevant user u is found at a rank position $r \leq N$, we say that a recommendation approach performed a *hit*; otherwise, we say it performed a *miss*. With only one relevant for each test instance, recall and precision at N can be formalized as presented in Figures 4.1 and 4.2.

In addition to recall and precision, since we are simulating a scenario with only one relevant result at a time for each new item, we also report mean reciprocal rank (MRR) [Baeza-Yates and Ribeiro-Neto, 2011], as:

$$\text{MRR} = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{r_j}, \tag{4.6}$$

[5] Once again, following standard practice, we use $\tau = 7$ for Book-Crossing and $\tau = 4$ for MovieLens-1M [Bellogin et al., 2011; Gunawardana and Meek, 2008].

$$\underbrace{\text{precision}(N) = \frac{\#\text{relevant retrieved}}{\#retrieved}}_{\text{Traditional precision}} \xrightarrow{\text{One relevant}} \frac{\#hits}{N \times \#tests} = \frac{\text{recall}(N)}{N}$$



precision(1) = 1 / (1*3) = 33%

precision(2) = 1 / (2*3) = 16%

precision(3) = 3 / (3*3) = 33%

Figure 4.2. Formal definition of precision at N with one relevant, followed by an example.

where r_j is the position of the relevant user in the ranked list produced for the j-th test instance and $|Q|$ is the number of test instances.

In particular, we perform a k-fold cross validation, i.e., for each of the two datasets described in Section 4.2, we randomly split the available items into k folds.[6] For each of k rounds, these folds are grouped into a training set, comprising $k-1$ folds, and a test set, comprising the remaining fold. Since we are simulating an extreme occurrence of the new item recommendation problem, if an item is in the test set, none of its ratings is used for training.

4.5 Parameters Setting

To train the parameter m of NIT-BR, which determines the maximum number of terms in the representation of each item (see Section 3.2), as well as the parameters of the several baselines used in our investigation (see Section 4.3), we perform a k-fold cross validation, optimizing for recall [Cremonesi et al., 2010]. In particular, Section 4.5.1 justifies the parameters for each variant of our model, while Section 4.5.2 does the same for the baselines.

4.5.1 NIT-BR model

In this section, we discuss the setting of the parameter m, which is the maximum number of terms to be selected, for different variants of our model (Section 3.2). These

[6] Following standard practice, we define $k = 5$ for Book-Crossing [Julià et al., 2009; Qumsiyeh and Ng, 2012] and $k = 10$ for MovieLens-1M [Cöster and Svensson, 2005; Schifanella et al., 2008].

variants are indexed according to the corresponding information-theoretic metric used as a term-weighting formula (see term $w(t, c)$ in Equation (3.2)). In particular, we limit the value of m to 30, otherwise, some items that have a description with fewer terms might end up with a complete representation.[7]

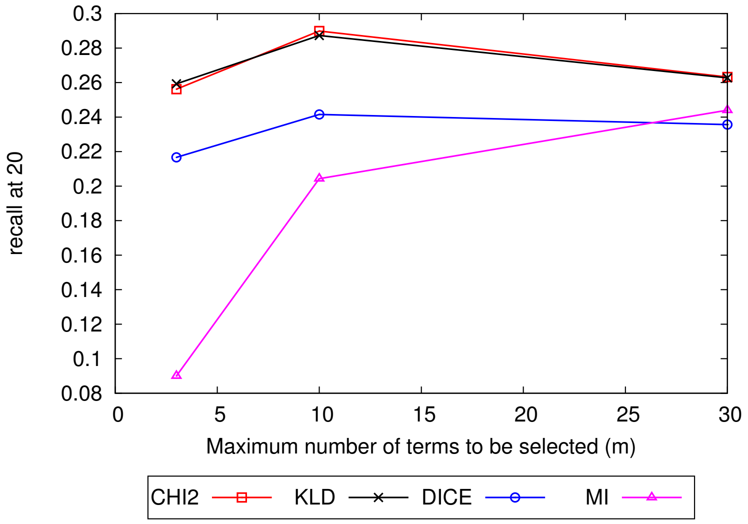


Figure 4.3. Recall@20 as a function of NIT-BR's parameter m for the Book-Crossing dataset.

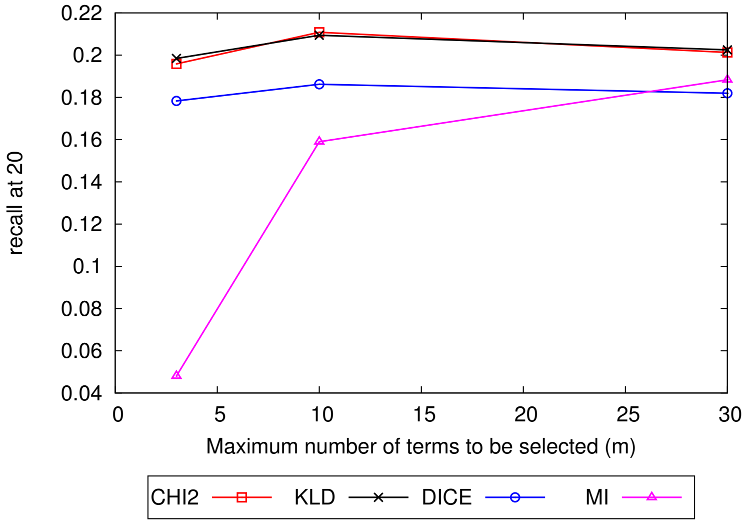


Figure 4.4. Recall@20 as a function of NIT-BR's parameter m for the MovieLens dataset.

[7] The NIT-BR$_{all}$ variant already reports the performance of our model using the full description of items as representation. As discussed in Section 4.3, this variant of our model is one of our baselines.

To support our investigation, Figures 4.3 and 4.4 report the recall@20 per m (maximum number of terms to be selected for the description of items) for all the aforementioned NIT-BR variants for both the Book-Crossing as well as the MovieLens-1M datasets. From Figures 4.3 and 4.4, we first observe that, ignoring an offset of the values, the performance of the variants reported behave similarly in both dataset, which suggests consistency. In addition, we notice that most metrics have the best performance with the value of the parameter around 10. The only exception is the Mutual Information (MI) variant, which presents the best performance with m = 30.

4.5.2 Baselines

In this section, we discuss the parameter settings for the baselines. In particular, we first discuss about tuning the number of latent factors for LSA. Next, we talk about tuning the taxonomy-weighting factor for LSA$_{tax}$.

LSA: Figure 4.5 reports LSA's performance in terms of recall@20 per number of latent factors for the Book-Crossing and MovieLens-1m datasets, respectively.[8] We observe that, in both datasets, the cut point of LSA performance is around 2,000 factors. In fact, we opt to use 2,000 factors since, with this parameterization, LSA's performance is close to the best, with a faster execution time.

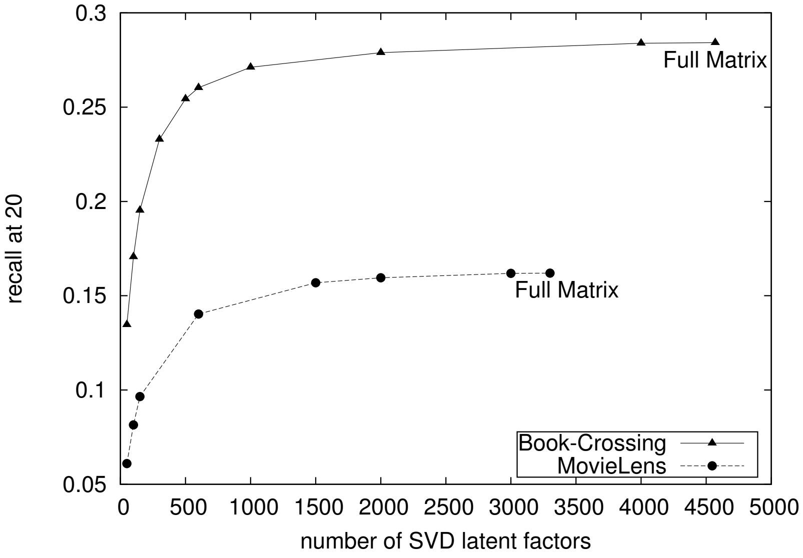


Figure 4.5. LSA's recall@20 per number of factors for both datasets.

[8] The "Full Matrix" entry, as suggested by its name, corresponds to the LSA's performance when using the entire matrix.

LSA$_{tax}$: Figure 4.6 reports the LSA$_{tax}$'s performance in terms of recall@20 per weight assigned to the taxonomic evidence for the Book-Crossing as well as the MovieLens-1m datasets. We observe that, in both datasets, the best performance is achieved by setting the taxonomy weight to 3.

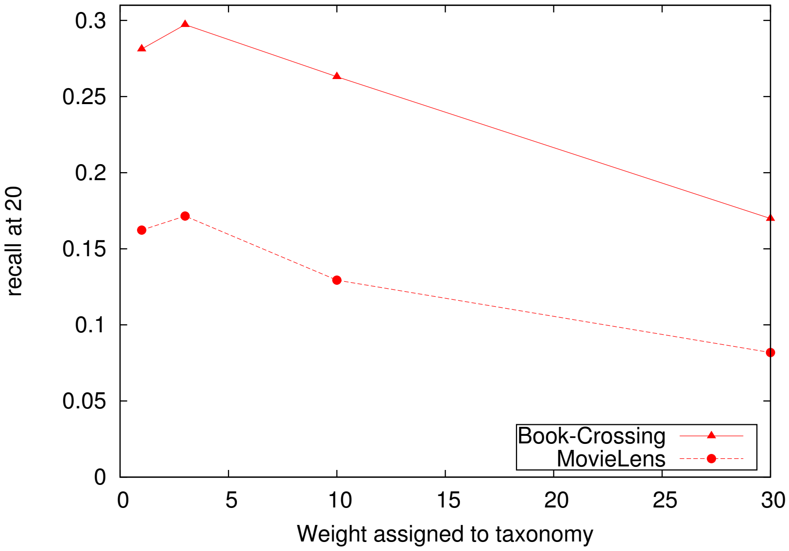


Figure 4.6. LSA$_{tax}$'s recall@20 per number of factors for both datasets.

From Figures 4.5 and 4.6, we observe that both baselines (LSA and LSA$_{tax}$, respectively) perform significantly better on the Book-Crossing dataset in comparison to their recommendation performance on the MovieLens-1m dataset. One possible explanation for this difference in recommendation performance is the discriminative power of the textual content of these two datasets. In particular, as illustrated in Table 4.1, Book-Crossing spans a much broader range of categories (855) compared to MovieLens-1m (18).

4.6 Summary

In this chapter, we have presented the experimental setup that supports the experimental evaluation presented in Chapter 5. First, we stated the research questions that we aim to answer. Next, we presented the recommendation datasets used in our experimentation, along with some standard pre-processing steps performed (such as the enrichment of these datasets with additional features). Then, we introduced our experimental methodology, which is a top-n recommendation task, similar to the one proposed by Cremonesi et al. [2010]. Finally, we defined our baselines and described how we tuned their parameters. In the next chapter, we will present our experimental evaluation.

Chapter 5

Experimental Evaluation

In this chapter, we thoroughly evaluate the effectiveness of our NIT-BR model at recommending new items. In order to answer the research questions stated in Chapter 4, we proceed as follows. In Section 5.1, we investigate the usefulness of taxonomies in the term selection component of our model, by contrasting the effectiveness of NIT-BR using the various term-category weighting schemes presented in Section 3.3. In Section 5.2, we assess the overall recommendation effectiveness of NIT-BR, by comparing it to state-of-the-art new item recommendation approaches from the literature, as described in Chapter 4. Finally, in Section 5.3, to evaluate the effectiveness of our model in domains where an explicit taxonomy is not available, we exploit automatically generated topics as a replacement for a taxonomy.

5.1 Taxonomy Usefulness (Q1)

In this experiment, we investigate the usefulness of taxonomies for improving the representation of items and users, therefore addressing our first research question. To this end, we assess the effectiveness of multiple variants of NIT-BR, each of which leveraging a different information theoretic metric (among those described in Section 3.3) in order to weigh the relative importance of each term with respect to the categories in the taxonomy underlying each of the two considered datasets. To assess the effectiveness of these variants (and, consequently, the usefulness of taxonomies), we contrast them with NIT-BR$_{all}$, a baseline variant of our model that does not perform any term selection. Instead, as defined in Equation (4.5), this variant represents an item with all the terms contained in the item description.

Figures 5.1 and 5.2 present recall(N) results for a range of rank cutoffs N for the Book-Crossing and MovieLens-1M datasets, respectively. From the figures, we first note that, in both datasets, the best results are attained by the KLD (Equation (3.6)) and CHI2 metrics (Equation (3.4)). Since these metrics perform similarly for most rank cutoffs, we combine them to create NIT-BR$_{comb}$, a variant of our model that estimates the relevance of a given new item to a user by linearly combining the KLD an CHI2 scores. This combination is statistically superior to all other metrics according to a paired t-test with $p < 0.05$. The only exception is a statistical tie with NIT-BR$_{chi2}$ in the 1st and 4th ranking positions for the Book-Crossing dataset.

Recalling our first research question, compared to the baseline variant NIT-BR$_{all}$, which does not perform term selection, most variants of NIT-BR improve, often significantly. In particular, for the Book-Crossing dataset, this baseline is significantly outperformed by all reported variants. For the MovieLens-1M dataset, the baseline is significantly outperformed by the Comb, CHI2, and KLD variants and statistically tied with the DICE and MI variants. Compared to our best-performing variant, NIT-BR$_{comb}$, the baseline is significantly outperformed in every ranking position, with improvements ranging from 31.86% to 50.86% for Book-Crossing and from 15.88% to 20.37% for MovieLens-1M. These results are corroborated by the MRR results in the middle part of Table 5.1, showing that NIT-BR$_{comb}$ significantly outperforms NIT-BR$_{all}$ by 40.49% and 16.03% in the Book-Crossing and MovieLens-1M datasets, respectively.

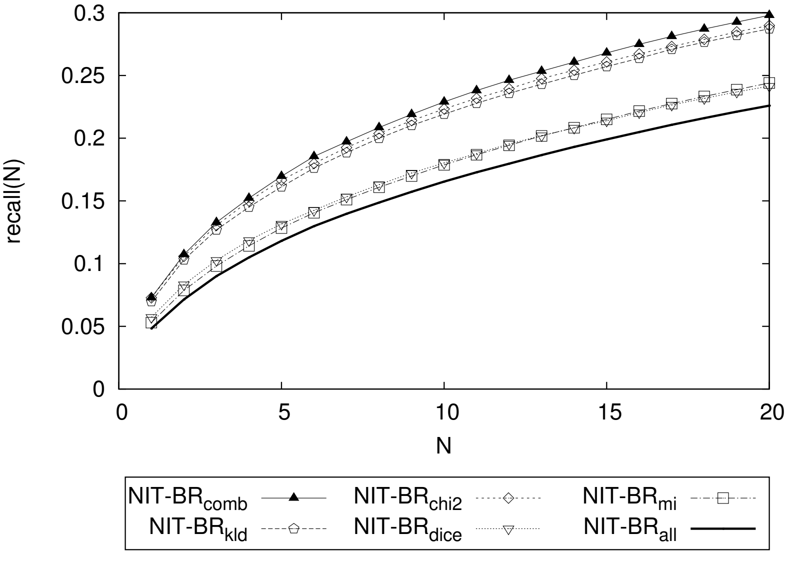


Figure 5.1. NIT-BR using various term selection strategies for the Book-Crossing dataset.

To further illustrate the benefits of our term selection variants, Table 5.2 presents the top terms selected by each variant to represent the movie “Titanic”, which belongs

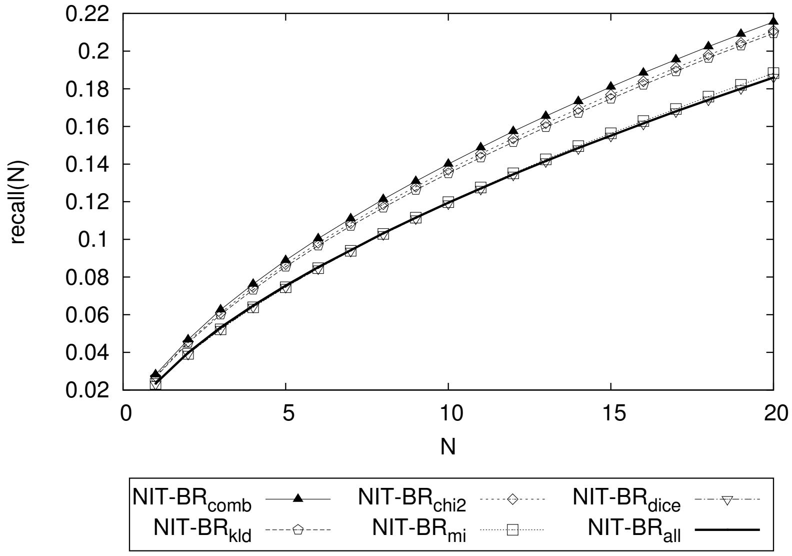


Figure 5.2. NIT-BR using various term selection strategies for the MovieLens-1M dataset.

Model	**Datasets**	
	Book-Crossing	MovieLens-1M
Random	0.0076 ± 0.0003	0.0075 ± 0.0001
TPU	0.0660 ± 0.0011	0.0507 ± 0.0012
STPU	0.0896 ± 0.0027	**0.0665** ± 0.0018
LSA	0.1088 ± 0.0061	0.0441 ± 0.0012
LSA$_{tax}$	**0.1148** ± 0.0060	0.0478 ± 0.0013
NIT-BR$_{all}$	0.0899 ± 0.0027	0.0599 ± 0.0016
NIT-BR$_{mi}$	0.0973 ± 0.0018	0.0599 ± 0.0013
NIT-BR$_{dice}$	0.1006 ± 0.0041	0.0596 ± 0.0016
NIT-BR$_{kld}$	0.1213 ± 0.0037	0.0672 ± 0.0020
NIT-BR$_{chi2}$	0.1241 ± 0.0039	0.0679 ± 0.0018
NIT-BR$_{comb}$	**0.1263** ± 0.0041	**0.0695** ± 0.0023
NIT-BR$_{lda}$	**0.1153** ± 0.0039	**0.0625** ± 0.0016
LSA$_{lda}$	0.1092 ± 0.0060	0.0441 ± 0.0012

Table 5.1. Mean reciprocal rank (MRR) for several recommendation approaches in the Book-Crossing and MovieLens-1M datasets.

to the categories *Drama* and *Romance* in the taxonomy of the MovieLens-1M dataset. From the table, we observe that the term selection mechanism of NIT-BR tends to discard terms that do not generalize well for other items that belong to the same categories as “Titanic”. For instance, while the term “collides” describes an important event in this particular movie, a user who likes the movie is arguably more interested in love stories than in maritime collisions.

Model	**Item Representation ($\hat{\imath}$)**
NIT-BR$_{all}$	existence passengers saved its suicide them his crash safe freezing making jack salvage and later titanic lifeboat of invited 1721 time memory the drawing ship collides her but strategy card ...
NIT-BR$_{mi}$	100yearold southampton 101yearold docksid 14th rms dine deepsea aft invalu forsak
NIT-BR$_{dice}$	love stori life jack friend woman name year tell one mother
NIT-BR$_{chi2}$	love fall marri woman jack fianc stori togeth name friend 1912
NIT-BR$_{kld}$	love woman stori life fall jack marri friend mother young togeth

Table 5.2. Example representations ($\hat{\imath}$) of the movie "Titanic" according to all variants of our NIT-BR model. To illustrate the representation of NIT-BR$_{all}$, which uses the entire description of an item, we randomly sampled 30 terms from a total of 526 that describe "Titanic". For the other variants, the terms are in decreasing order of importance.

Overall, the results in this section answer our first research question, by showing that exploiting taxonomies to aid the selection of informative terms brings significant improvements to the effectiveness of our model. NIT-BR$_{comb}$, the variant of our model that combines the scores of the KLD and CHI2 metrics, is particularly effective, substantially and significantly improving upon the use of all terms from the description of the items. Furthermore, the consistent improvements obtained by the various considered information theoretic metrics attest the robustness of the taxonomic evidence when exploited to improve the representation of items and users in different datasets.

5.2 NIT-BR Effectiveness (Q2)

In the previous section, we showed the positive impact of exploiting taxonomies to aid the selection of informative terms for an improved representation of items and users within our NIT-BR model. In this section, we compare the best variant of our model, NIT-BR$_{comb}$, with state-of-the-art new item recommendation baselines from the literature, namely, TPU (Equation (4.1)), STPU (Equation (4.2)), LSA (Equation (4.3)), and LSA$_{tax}$ (Equation (4.4)). Figures 5.3, 5.4, 5.5 and 5.6 show the results of this investigation for the Book-Crossing and MovieLens-1M datasets, respectively, in terms of recall(N) and precision vs. recall curves. In addition, the top and middle parts of Table 5.1 provide MRR figures.

From Figures 5.3 and 5.5, we first notice that our NIT-BR$_{comb}$ variant, which exploits taxonomies to select informative terms using a combination of the CHI2 and KLD

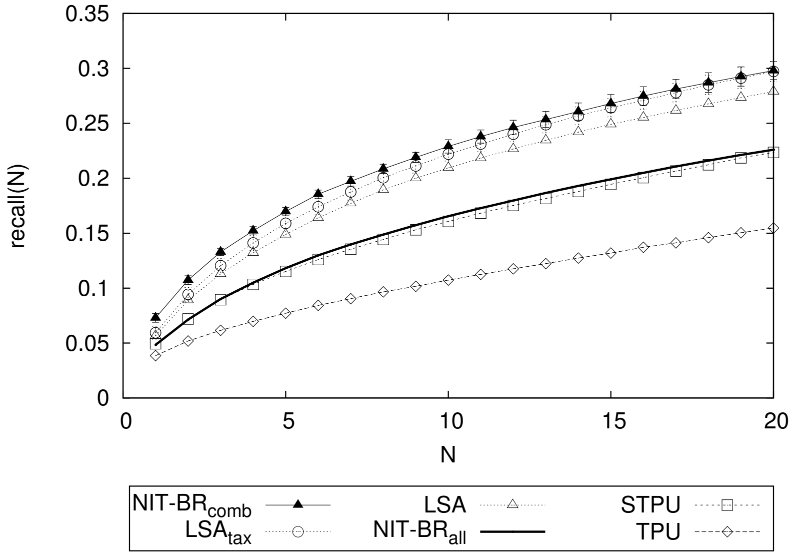


Figure 5.3. Recall at N for NIT-BR and baseline recommenders for the Book-Crossing dataset.

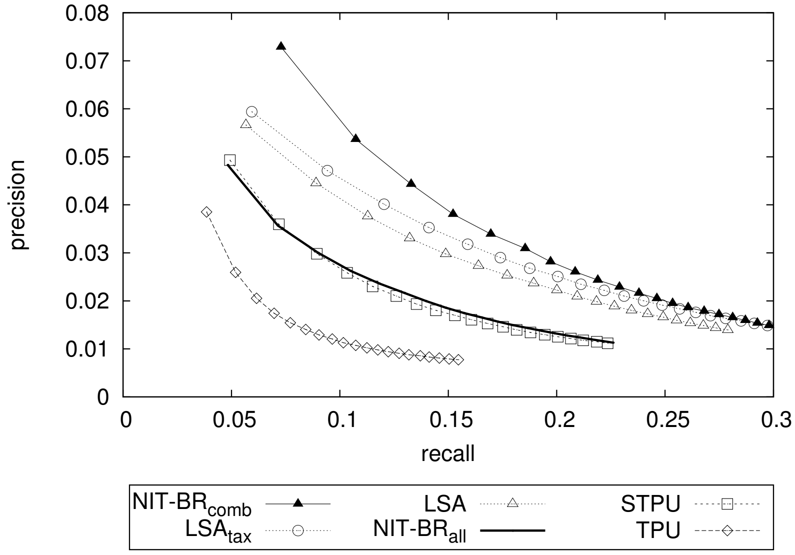


Figure 5.4. Precision vs. recall for NIT-BR and baseline recommenders for the Book-Crossing dataset.

information theoretic metrics, has the best recall(N) among all considered approaches, with significant improvements of up to 22.68% and 6.87% in the Book-Crossing and MovieLens-1M datasets, respectively. In addition, Figures 5.4 and 5.6 show that, for both datasets, NIT-BR$_{comb}$ significantly outperforms all baselines also in terms of precision, with gains at all recall levels. This is exactly what we want, to be able to

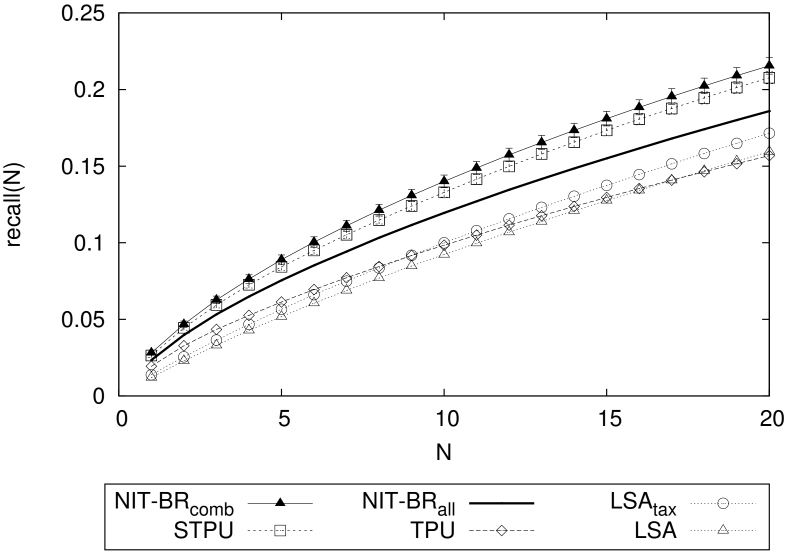


Figure 5.5. Recall at N for NIT-BR and baseline recommenders for the MovieLens-1M dataset.

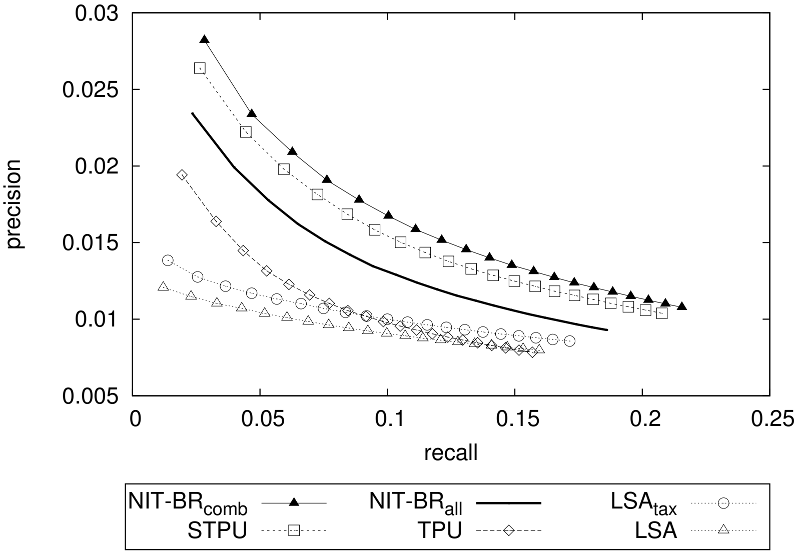


Figure 5.6. Precision vs. recall for NIT-BR and baseline recommenders for the MovieLens-1M dataset.

find a small set of users who will be interested in a given new item, without incurring too many false positive recommendations. In terms of MRR, as shown in Table 5.1, NIT-BR$_{comb}$ significantly outperforms the strongest baseline in each dataset, with gains of 10.01% over LSA$_{tax}$ for Book-Crossing, and 4.51% over STPU for MovieLens-1M. Finally, compared with LSA$_{tax}$, which also exploits taxonomies to improve upon the pure content-based LSA approach, the results in Figures 5.3, 5.4, 5.5 and 5.6 and also Table 5.1 attest the effectiveness of our simple information-theoretic approach for identifying informative terms to describe the cataloged items.

Overall, the results reported in this section reinforce our observations in the previous section regarding the usefulness of taxonomies to improve the representation of items and users in a recommender system. Moreover, they attest the effectiveness of our NIT-BR model in contrast to state-of-the-art approaches from the literature for the new item recommendation problem, hence answering our second research question. In the next section, we investigate the impact of using automatically generated categories when no explicit taxonomy is available for a particular domain.

5.3 Automatically Generated Categories (Q3)

In Sections 5.1 and 5.2, we validated the usefulness of taxonomies as well as the overall effectiveness of our NIT-BR model for the new item recommendation problem, respectively. In this section, we aim to answer our third and last research question, regarding the impact of automatically generated categories as a replacement of taxonomy categories for domains where an explicit taxonomy is not available.

To address our third research question, we deploy the best variant of our model, NIT-BR$_{comb}$, using either an explicit taxonomy, as in the previous sections, or an automatically generated taxonomy, with categories represented by topics automatically identified using Latent Dirichlet Allocation (LDA) Blei et al. [2003]. For the sake of clarity, we refer to the latter as NIT-BR$_{lda}$. To assess the effectiveness of our approach, we compare our results with LSA$_{lda}$, an extended version of the LSA algorithm that includes the LDA's topics into the model.[1] In addition, we once again include NIT-BR$_{all}$, the baseline variant of our model, which performs no term selection. Figures 5.7 and 5.8 presents recall(N) results for the aforementioned recommendation approaches for both the Book-Crossing as well as the MovieLens-1M datasets.

From Figures 5.7 and 5.8, we observe that NIT-BR$_{lda}$ significantly outperforms NIT-BR$_{all}$, with gains in recall ranging from 22.41% to 35.82% for Book-Crossing, and from 3.71% to 5.34% for MovieLens-1M. While leveraging an explicit taxonomy can further improve, as observed from the performance of NIT-BR$_{comb}$, this result shows the feasibility of deploying our model for domains where such a taxonomy is not available. In addition, compared to LSA$_{lda}$, NIT-BR$_{lda}$ is significantly superior at the

[1] LSA$_{lda}$ is similar to LSA$_{tax}$, except that the former leverages latent topics instead of explicit categories.

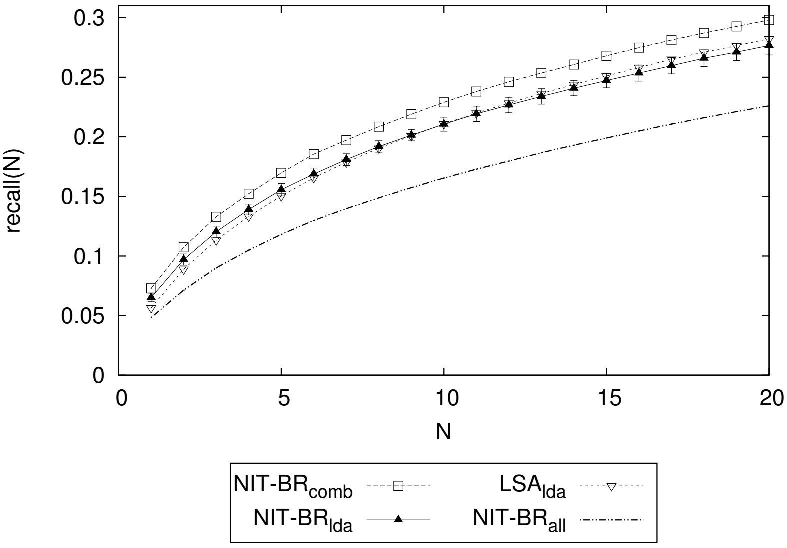


Figure 5.7. NIT-BR using explicit vs. automatically generated categories for the Book-Crossing dataset.

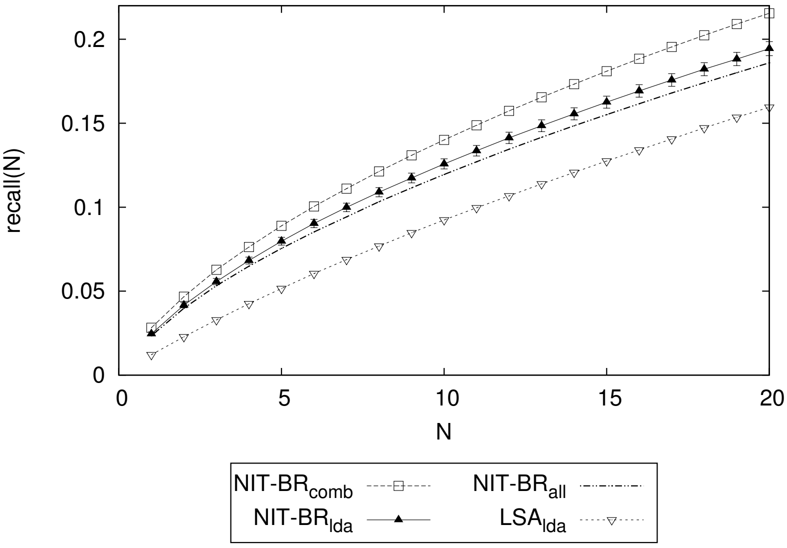


Figure 5.8. NIT-BR using explicit vs. automatically generated categories for the MovieLens-1M dataset.

top six ranking positions for Book-Crossing (Figure 5.7), with improvements ranging from 15.13% to 1.90%. For lower ranking positions, NIT-BR$_{lda}$ is statistically tied in positions 6 to 13, and statistically inferior in the remaining positions. For MovieLens-1M, NIT-BR$_{lda}$ significantly outperforms LSA$_{lda}$ in every ranking position, with gains in recall(N) ranging from 101.72% to 21.87%. These observations are also consistent in terms of MRR, as shown at the bottom of Table 5.1.

Recalling our third research question, the results in this section attest the effectiveness of NIT-BR compared to a state-of-the-art content-based recommendation baseline from the literature, even when no explicit taxonomy is available. On the other hand, while categories automatically derived using LDA can be used effectively by our model, the results in this section also show that the availability of an explicit, manually curated taxonomy can provide further gains.

5.4 Summary

In this chapter, we showed experimental results that answered our research questions and evaluated different aspects of our model. First, we discussed the effects of using a taxonomy within our model, as our first experimentation showed that the selection of taxonomy-like terms using information-theoretic metrics significantly improves upon the use of all terms from the description of the items. Next, we compared the best variant of our model, namely NIT-BR$_{comb}$, to state-of-the-art baselines from the literature, especially LSA$_{tax}$, the most directly comparable baseline. Again, results showed the effectiveness of our model in selecting users interested in a given new item. Finally, our last research question was answered by showing that our model still performs effectively even when no explicitly taxonomy is available. In the next chapter, we expose our conclusions and plans for future work.

Chapter 6

Conclusions and Future Work

In this dissertation, we introduced New Item Taxonomy-Based Recommender (NIT-BR), a novel approach for new item recommendation, a challenging problem for current recommender systems that must cope with continuously evolving item catalogs. In particular, NIT-BR tackles this problem as a classical search problem, by modeling the terms that describe the new item as a "query", and each candidate user who could be recommended the item as a "document", comprising the terms in the description of the items that the user has positively rated in the past. To improve this content-based representation, we proposed a term selection mechanism aimed to weigh the informativeness of each term with respect to the taxonomy categories covered by each item, based upon classical information-theoretic metrics.

We thoroughly investigated the effectiveness of our NIT-BR model at recommending new items. By contrasting our model with a variant that performs no term selection, we demonstrated the usefulness of taxonomies as a source of evidence for improving the underlying content-based representation of items and users. In particular, this improved representation was shown to consistently and significantly outperform state-of-the-art recommendation baselines from the literature across two publicly available datasets covering distinct domains, namely, book and movie recommendations. Lastly, we demonstrated the feasibility of leveraging automatically generated categories based on topic modeling for domains where an explicit taxonomy is not available.

Besides being effective in practice, our model can also produce more interpretable recommendations, allowing a recommender system to better explain the recommendations provided for a given user. For instance, suppose the movies "Titanic" and "Romeo and Juliet" share the same genre *Drama*, and that a given user has positively rated the first movie. A recommender system that uses NIT-BR can inform to the user that "Romeo and Juliet" was suggested as a recommendation due to the positive rating the user gave to "Titanic".

As pointed out in Section 3.1, in Equation (3.1), we opted for a simple yet effective TF-IDF formulation operating on top of the term-based representations of items and users. In the future, we plan to further investigate the performance of our model using alternative similarity functions, such as BM25 [Robertson et al., 2004] and Language Modeling [Ponte and Croft, 1998]. Moreover, as discussed in Section 5.1, our best reported variant is a linear combination of two other variants of our model. As future work, we aim at exploring machine learning techniques, e.g., learning-to-rank, to achieve a better effectiveness while combining these variants. Finally, we also intend to apply our model to the new user recommendation problem presented in Section 2.2.2. In particular, for the non-extreme version of the problem, we can exploit a few ratings associated with the user to drive his or her content-based representation.

Bibliography

Adomavicius, G. and 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):734--749.

Ahmed, A., Kanagal, B., Pandey, S., Josifovski, V., Pueyo, L. G., and Yuan, J. (2013). Latent factor models with additive and hierarchically-smoothed user preferences. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining*, pages 385--394.

Anand, S. S. and Griffiths, N. (2011). A market-based approach to address the new item problem. In *Proceedings of the 5th ACM Conference on Recommender Systems*, pages 205--212.

Baeza-Yates, R. A. and Ribeiro-Neto, B. A. (2011). *Modern Information Retrieval - the concepts and technology behind search.* Pearson Education Ltd., second edition.

Bambini, R., Cremonesi, P., and Turrin, R. (2011). A recommender system for an iptv service provider: a real large-scale production environment. In Ricci, F., Rokach, L., Shapira, B., and Kantor, P. B., editors, *Recommender Systems Handbook*, pages 299–331. Springer US.

Bellogin, A., Castells, P., and Cantador, I. (2011). Precision-oriented evaluation of recommender systems: An algorithmic comparison. In *Proceedings of the 5th ACM Conference on Recommender Systems*, pages 333--336.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993--1022.

Brandão, W. C., de Moura, E. S., Santos, R. L. T., da Silva, A. S., and Ziviani, N. (2013). Learning to Expand Queries Using Entities. *Journal of the American Society for Information Science and Technology.*

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331--370.

Cho, Y. H. and Kim, J. K. (2004). Application of web usage mining and product taxonomy to collaborative recommendations in e-commerce. *Expert Systems with Applications*, 26(2):233--246.

Cöster, R. and Svensson, M. (2005). Incremental collaborative filtering for mobile devices. In *Proceedings of the 20th ACM Symposium on Applied Computing*, pages 1102--1106.

Cremonesi, P., Koren, Y., and Turrin, R. (2010). Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the 4th ACM conference on Recommender systems*, pages 39--46.

Cremonesi, P. and Turrin, R. (2009). Analysis of cold-start recommendations in iptv systems. In *Proceedings of the 3rd ACM Conference on Recommender Systems*, pages 233--236.

Cremonesi, P., Turrin, R., and Airoldi, F. (2011). Hybrid algorithms for recommending new items. In *Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, pages 33--40.

Deshpande, M. and Karypis, G. (2004). Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems*, 22(1):143--177.

Furnas, G. W., Deerwester, S., Dumais, S. T., Landauer, T. K., Harshman, R. A., Streeter, L. A., and Lochbaum, K. E. (1988). Information retrieval using a singular value decomposition model of latent semantic structure. In *Proceedings of the 11th ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 465--480.

Gedikli, F. and Jannach, D. (2010). Recommending based on rating frequencies. In *Proceedings of the 4th ACM Conference on Recommender Systems*, pages 233--236.

Gunawardana, A. and Meek, C. (2008). Tied boltzmann machines for cold start recommendations. In *Proceedings of the 2nd ACM Conference on Recommender Systems*, pages 19--26.

Julià, C., Sappa, A., Lumbreras, F., Serrat, J., and López, A. (2009). Predicting missing ratings in recommender systems: Adapted factorization approach. *International Journal of Electronic Commerce*, 14(2):89--108.

Jurgens, D. and Stevens, K. (2010). The S-Space package: an open source package for word space models. In *Proceedings of the 48th ACL System Demonstrations*, pages 30--35.

Kanagal, B., Ahmed, A., Pandey, S., Josifovski, V., Yuan, J., and Garcia-Pueyo, L. (2012). Supercharging recommender systems using taxonomies for learning user purchase behavior. *Proceedings of the VLDB Endowment*, 5(10):956--967.

Lam, X. N., Vu, T., Le, T. D., and Duong, A. D. (2008). Addressing cold-start problem in recommendation systems. In *Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication*, pages 208--211.

Leung, C. W.-k., Chan, S. C.-f., and Chung, F.-l. (2008). An empirical study of a cross-level association rule mining approach to cold-start recommendations. *Knowledge-Based Systems*, 21(7):515--529.

Levi, A., Mokryn, O., Diot, C., and Taft, N. (2012). Finding a needle in a haystack of reviews: cold start context-based hotel recommender system. In *Proceedings of the 6t ACM Conference on Recommender Systems*, pages 115--122.

Liu, N. N., Meng, X., Liu, C., and Yang, Q. (2011). Wisdom of the better few: cold start recommendation via representative based rating elicitation. In *Proceedings of the 5t ACM Conference on Recommender Systems*, pages 37--44.

Liu, Y., Loh, H. T., and Sun, A. (2009). Imbalanced text classification: A term weighting approach. *Expert Systems with Applications*, 36(1):690--701.

Lops, P., Gemmis, M., and Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. In *Recommender Systems Handbook*, pages 73--105. Springer.

Mahmood, T. and Ricci, F. (2009). Improving recommender systems with adaptive conversational strategies. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, pages 73--82.

Miller, B. N., Albert, I., Lam, S. K., Konstan, J. A., and Riedl, J. (2003). Movielens unplugged: experiences with an occasionally connected recommender system. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, pages 263--266.

Park, S.-T. and Chu, W. (2009). Pairwise preference regression for cold-start recommendation. In *Proceedings of the 3rd ACM Conference on Recommender Systems*, pages 21--28.

Park, S.-T., Pennock, D., Madani, O., Good, N., and DeCoste, D. (2006). Naïve filterbots for robust cold-start recommendations. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 699--705.

Pilászy, I. and Tikk, D. (2009). Recommending new movies: even a few ratings are more valuable than metadata. In *Proceedings of the 3rd ACM Conference on Recommender Systems*, pages 93--100.

Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275--281.

Qumsiyeh, R. and Ng, Y.-K. (2012). Predicting the ratings of multimedia items for making personalized recommendations. In *Proceedings of the 35th ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 475--484.

Resnick, P. and Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, 40(3):56--58.

Ricci, F., Rokach, L., and Shapira, B. (2011). Introduction to recommender systems handbook. In *Recommender Systems Handbook*, pages 1–35. Springer.

Robertson, S., Zaragoza, H., and Taylor, M. (2004). Simple BM25 extension to multiple weighted fields. In *Proceedings of the 13th ACM International Conference on Information and Knowledge Management*, pages 42--49.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513--523.

Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2000). Analysis of recommendation algorithms for e-commerce. In *Proceedings of the 2nd ACM Conference on Electronic Commerce*, pages 158--167.

Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, pages 285--295.

Sarwar, B. M., Karypis, G., Konstan, J., and Riedl, J. (2002). Incremental svd-based algorithms for highly scaleable recommender systems. In *Proceedings of the 5th International Conference on Computer and Information Technology*, pages 27--28.

Schafer, J. B., Konstan, J., and Riedi, J. (1999). Recommender systems in e-commerce. In *Proceedings of the 1st ACM Conference on Electronic Commerce*, pages 158--166.

Schein, A. I., Popescul, A., Ungar, L. H., and Pennock, D. M. (2002). Methods and metrics for cold-start recommendations. In *Proceedings of the 25th ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 253--260.

Schifanella, R., Panisson, A., Gena, C., and Ruffo, G. (2008). Mobhinter: epidemic collaborative filtering and self-organization in mobile ad-hoc networks. In *Proceedings of the 2nd ACM Conference on Recommender Systems*, pages 27--34.

Weng, L.-T., Xu, Y., Li, Y., and Nayak, R. (2008). Exploiting item taxonomy for solving cold-start problem in recommendation making. In *Proceedings of the 20th IEEE International Conference on Tools with Artificial Intelligence*, pages 113--120.

Zhang, W. (2008). Relational distance-based collaborative filtering. In *Proceedings of the 31st ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 877--878.

Zhou, K., Yang, S.-H., and Zha, H. (2011). Functional matrix factorizations for cold-start recommendation. In *Proceedings of the 34th ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 315--324.

Ziegler, C.-N., Lausen, G., and Schmidt-Thieme, L. (2004). Taxonomy-driven computation of product recommendations. In *Proceedings of the 13th ACM Conference on Information and Knowledge Management*, pages 406--415.

Ziegler, C.-N., McNee, S. M., Konstan, J. A., and Lausen, G. (2005). Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web*, pages 22--32.