文章编号：1003-0077（2011）00-0000-00

基于简介和评论的标签推荐方法研究 *

褚晓敏，王中卿，朱巧明，周国栋

（苏州大学 计算机科学与技术学院，江苏 苏州 215006）

摘要： Web 2.0 时代，社会标签是信息资源组织的一种重要方式。标签推荐能够有效的帮助用户收集，定位，查找和共享在线资源。以往的标签推荐算法只是基于一种文本信息，比如基于电影的简介文本来进行标签推荐。但是实际上电影往往存在多种文本信息，比如同时存在摘要信息和评论信息，不同类型的信息能够反映电影的不同方面的属性，因此为了提高电影标签推荐的准确率和有效性，我们同时根据电影的简介和短评进行电影标签自动推荐，并使用多种方法融合基于不同类型文本的标签推荐的结果，实验证明，使用不同类型信息进行标签推荐能够比单一使用一种文本信息进行标签推荐有很大的提升

关键词：自然语言处理；社会标签；社会关系网络；分类器融合

中图分类号：TP391　　**文献标识码：**A

Tag Recommendation with Summary and Comment Information

CHU Xiaomin, WANG Zqing, ZHU Qiaoming, ZHOU Guodong

(School of Computer Sciences and Technology, Soochow University, Suzhou, Jiangsu 215006, China)

Abstract: Social tags are important styles of information organizing on the Web 2.0 era. Tag recommendation can help users collect, search and share online resources effectively. The previous approaches focus on using single types of textual information, e.g. summary of a movie. But in practice there are various types of textual information can be used for tag recommendation, for example, a movie contains both summary and comment information. Different types of information reflect the different aspects of the movie. Thus we propose a novel approach to combine both summary and comment information to recommend tags; furthermore, we use different ensemble learning approaches to incorporate the above information. The experimental results show that our proposed approach with using different types of information can outperform using single types of textual information in the tag recommendation tasks.

Key words: natural language processing; social tags; ensemble learning

1 引言

Web 2.0 时代，人们很容易为各种在线资源标注标签，由此诞生了众多的标签推荐系统，如 Folksonomy、Delicious 等。社会标签作为信息资源的组织式，越来越受到网络用户的欢迎，人们已经习惯于使用标签来定位、收集和共享在线资源，例如网页、照片、视频、电影、书籍等。

并不是所有网站都提供针对实体资源的标签或标签推荐，因此自动标签推荐是一个十分重要的任务，通常自动标签推荐是指通过考察、分析、挖掘信息资源的内容和用户的历史标注以及显式或隐式的关系为未标注信息资源提供高质量的候选标签。标签推荐的目的是：1）简化标注活动，为用户提供方便，并增加标签的可用性和粘性；2）提高标签质量，降低错拼、歧义等情况，提高标签在信息资源组织、检索、利用和发现中的作用；3）改变标签空间的结构，使标签空间更快的稳定和收敛。

通常标签推荐只是使用一种信息进行推荐，最常用的就是各类简介，相对于评论、讨论

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金（61272260, 61331011）；江苏省高校自然科学基金重大项目（11KJA520003）

等信息，简介是对实体或资源内容或属性的描述，一般来说更客观。但事实上，实体或资源，比如电影，存在多种类型的信息，如短评、影评、问答、讨论等，每一类信息都能反映电影一方面的信息，如短评、影评等能够反映用户对一部电影的主观评价，因此结合多方面的信息进行标签推荐能够获得更好的推荐效果。

在本文中，我们主要研究利用电影的剧情简介和短评信息进行标签推荐，并基于各种信息的推荐结果进行多种方法的融合，以提高推荐效果。实验结果表明利用不同种类的信息能够有效的提高标签推荐的准确率和有效性。

以下各章节组织如下：第一章节介绍标签推荐的相关工作，第二章节介绍本文的数据收集和统计情况，第三章节介绍本文的研究内容和使用的研究方法，第四章节介绍实验设置和相关结果，第五章节总结本文内容以及对下一步工作做初步计划。

2 相关工作

在标签推荐系统中，任何用户都具有开放参与的特性，均可用自定义标签标注资源；在铺天盖地的信息资源中用户可以通过标签更方便和准确的定位、查找和共享资源。标签推荐系统根据资源内容、用户信息、已有标签，自动提供一些与资源内容相关或者用户感兴趣的标签供用户选择，大大减少用户标注所需时间，进而改善用户体验。目前，社会标签已经广泛的被研究，并成功的应用在标签推荐 [1-3]、趋势预测和跟踪 [4]、个性化 [5]、广告推送 [6] 等系统中。标签推荐的方法可以分为两大类，即基于内容（content-based）的标签推荐和基于协同过滤（CF-based）的标签推荐。这两类方法各自利用了标签推荐问题中不同来源的信息来完成标签推荐任务。

2．1 基于内容的标签推荐

基于内容的方法是标签推荐的基本方法，往往与其他方法联合使用。基于内容的标签推荐以文档的内容为标签推荐的依据，可使用文档的细粒度特征和粗粒度特征这两种主要方式。1）使用文档的细粒度特征，如词汇。例如，Ohkura et al.[7]用 SVM 为每个标签训练了一个文本分类器，根据文档内容计算哪个标签更合理，其研究成果已经应用于基于社会标签的网页浏览辅助系统中。Mishne et al.[8]利用近邻法（KNN）进行标签推荐，从文档集合中选择与新文档最相关的 K 个文档，将这 K 个文档的标签推荐给新文档。2）使用文档的粗粒度特征，如主题。此时不再考虑单个词汇与标签之间的关系，而是通过主题模型 LDA（Latent Dirichlet Allocation）[9-10]，抽取新的文档与已有标签的主题特征，找出其相似度，根据相似度推荐标签。例如，Blei et al. [11]提出了一种有监督的主题模型，这种模型是对 LDA 的一种改进，增加了一个连续变量代表标签，并在此模型上训练出最优的参数。Si et al. [12]在 LDA 模型基础上提出 Tag-LDA，基于文档内容和标签联合建模，并取得了比较好的推荐效果。

2．2 基于链接的标签推荐

目前研究最多的标签推荐技术是基于协同过滤（CF）[13-14]的技术，该方法根据用户群中其他相似用户的兴趣和爱好推断用户可能需要的资源，利用该方法建立标签推荐系统已经成为一种常见的研究方法 [15]。基于协同过滤的典型做法是基于给定的资源和用户的标签历史进行标签推荐。例如，Nakamoto et al.[16]依据标签推荐系统中用户定义标签习惯的相似程度进行基于模式的协同过滤，通过用户聚类的方法推荐标签。Niwa et al.[17]、Gemmell et al.[18] 借鉴了 TF-IDF 算法，分别依据标签与资源的紧密度和利用层次聚类法进行标签聚类。Santos-Neto et al. [19]通过构建用户网络网，按结构寻找相似团体进行协同过滤推荐。Liu et al. [20]提出基于连续条件随机场的标签推荐模型进行标签推荐，在保证条件概率最大的情况下通过训练得出模型参数，在执行模块中得出排名分数前十的标签。FlokRank[21]和矩阵分解 [22] 是基于 CF 方法进行社会标签推荐的代表性方法。这些方法最常见的是冷启动问题，也就是说如果没有被标注过，就很难进行有效的标签推荐。

本文使用基于内容的标签推荐方法进行电影标签的研究，在上述的基于内容的标签推荐方法中，通常只使用了一方面的文本信息，而本文使用了两类文本信息来进行标签推荐，以获得更好的标签推荐效果。

3 数据统计

3．1 数据收集与统计

本文使用的数据来源于豆瓣网。从互联网获取“豆瓣电影”上的电影信息，使用爬虫工具抓取 1751 个电影的数据。通过数据预处理，选择其中标签、剧情简介、短评都齐全的 1634 部电影数据，抽取这些电影的标签、剧情简介和前 20 条短评数据进行本文相关实验。

表 1　数据集中包含标签的数据量

Table 1　Number of tags on data set

标签	标签数量（个）	标签	标签数量（个）
美国	899	科幻	368
喜剧	479	经典	322
爱情	450	动作	273
剧情	371	搞笑	206
动画	369	香港	159

在我们收集的电影数据中，一共存在 2204 个不同的标签，但是由于大部分标签出现频率很低并有部分重复的现象，因此我们选择使用频度最高并且不重复的 10 个标签，美国、喜剧、爱情、剧情、动画、科幻、经典、动作、搞笑、香港，并分别使用剧情简介、短评作为特征内容进行分类器训练。这 10 个标签在数据集中包含标签的数据量如表 1 所示。

3．2 数据样例

表 2 给出了一个具体电影的样例，包含三方面信息，剧情简介、标签（标签后面括号内的数字为标记此种标签的用户数）和短评。从样例中我们可以看出，标签简明的指明了资源的主要内容、特点以及用户的兴趣点。针对样例《被解救的姜戈》，标签“黑色幽默”表明了电影的主要特点，标签“美国”表明了电影的发行方或故事的主要发生地点，标签“西部”、“暴力”、“动作”则让观众了解到该部电影的主要题材和类型。剧情简介摘要的说明了电影的发生背景和主要内容，以描述性文字为主，短评则是针对电影的评论，评论可能是针对内容的，针对编剧或导演的，针对电影演员的，针对场面和特技的，甚至是纯粹的吐槽，以评论性文字为主。所列举的标签中，“美国”、“暴力”、“动作”这些标签可以从剧情简介中获得，而“暴力”、“西部”、 “西部片”则可以从电影的短评中获得相关信息。因此剧情简介和短评体现了不同类型的信息，并且都可以作为标签推荐的基础。

表 2　电影信息样例

Table 2　Sample of a movie

电影名称：被解救的姜戈
剧情简介： 1858 年，美国南北战争前两年。德国赏金猎人金·舒尔茨从贩奴商人手中买下黑奴姜戈，……本片是昆汀向 1966 年由塞吉奥·考布西执导的经典意大利西部片《迪亚戈》的致敬之作……本片延续昆汀一贯低调奢华的风格，处处可见奇思妙想的幽默元素和血腥野性的动作场面……
短评： ♦ 或许不是年度最佳电影，但极可能是最令影迷们过瘾的电影……混搭音乐、暴力、语言（特别是外语）再一次成为手下重要的道具…… ♦ 一部很 high 的意大利通心粉西部片，就是那种影迷们期待的痞子昆电影…… ♦ 昆汀的暴力美学总是这么精彩 ♦ 我咋就是瞧不上西部片呢。。。 ♦ 前抑后扬，情绪逐级酝酿……老式西部片通常都是杀人不见血，一声枪响一股烟，人就躺下了，昆汀这已经不是血浆了，是血雨，爆开的鲜血之雨。而鲜血染白花也成为昆汀拿手的暴力审美……
标签：美国(26185) 西部(20855) 黑色幽默(14960) 暴力(13191) 剧情(6213) 2012(6068) 动作(5759) 西部片(5155)

4 基于简介和评论的电影标签推荐

为了融合简介和评论两方面的信息进行电影标签推荐，我们将推荐任务转化为分类任务，抽取简介和评论文本的单词作为特征，使用 SVM 构建基分类器，并使用不同的方法进

行分类器融合。整体研究框架主要步骤包括：1）从互联网上获取批量的电影数据；2）分析网页获取标签，简介文本，评论文本等信息；3）对原始数据进行预处理，包括数据选择、分词等；4）训练剧情简介、短评这两个基分类器；5）进行分类器融合，采用直接融合、投票规则和加法规则等策略；6）分析比较实验结果。具体的研究框架和流程如图 1 所示：

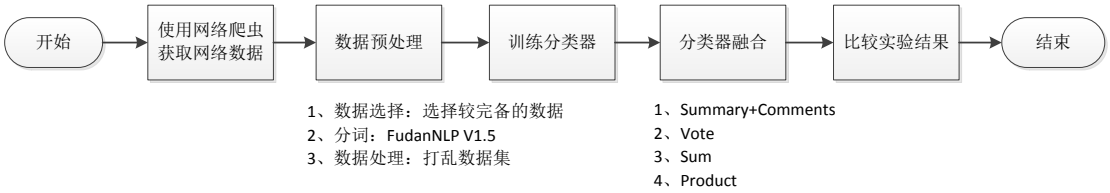

图 **1**　整体研究框架

Fig. 1　Framework of our proposed approach

对于数据预处理的过程，我们选择了剧情简介和短评信息都完备的数据样本，使用复旦的分词工具 [1]进行分词处理，做好数据准备和预处理后再进行分类器的训练和融合。

我们把推荐问题转化为分类问题，为每个标签建立一个分类器，根据包含标签的数量进行正负样本数据的选择和整理，如果样本包含标签则归入正类样本，不包含则归入负类样本，因为正类样本数量较少，我们通过欠采样策略使样本保持平衡。

为了融合基于不同信息的基分类器的结果，我们提出了多种策略进行融合，具体包括直接融合、投票规则、加法规则等 [23]。为了描述我们的分类器融合算法，我们先进行如下形式化定义：

R 是分类器 $c_k\left(k=1,...,R\right)$ 中的元素，每个分类器作为样本输入（以 x_k 来表示）标签 L_k（$L_k=w_1,\ldots,w_m$）。假设分类器 C_k 的输出度量值，代表后验概率向量，$P_k=\left[p\left(w_1\mid x_k\right),\ldots,p\left(w_m\mid x_k\right)\right]^t$，（ $\left(p\left(w_i\mid x_k\right)\right.$ 表示 x 标记为 w_i 的概率）。

1） 直接融合

直接融合时，我们将剧情简介和短评这两方面文本信息进行文本组合后作为一个特征送入分类器，进行分类训练，其分类结果也作为基分类器参与其他融合方法。

2） 投票规则

将简介、短评、直接融合三种方法训练出来的基分类器作为输入，包含标签时投 1 票，不包含标签时投 0 票，投票结果大于等于 2 则表示测试样本包含测试标签，否则不包含测试标签。应用投票（Vote）规则时：

$$assign\ Z\rightarrow w_j \qquad j=\arg\max_i\sum_{i=1}^{R}\Delta_i \qquad \text{其中，}\ \Delta_i=\begin{cases}1 & L_k=w_i\\ 0 & L_k\neq w_i\end{cases}$$

3） 加法规则

将简介、短评、直接融合三种方法训练出来的基分类器作为输入，正值概率和负值概率分别相加，如果正值概率和>负值概率和，则测试样本包含测试标签，否则不包含测试标签。

$$assign\ Z\rightarrow w_j \qquad j=\arg\max_i\left\{p(w_i)\sum_{k=1}^{R}p(w_i\mid x_k)\right\}$$

5　实验

5．1 实验设置

本实验使用的数据来源于豆瓣电影的电影信息，使用爬虫工具从互联网获取。选择使用频度最高的 10 个标签，分别使用剧情简介、短评作为特征内容进行训练。对抽取的 1751 个电影数据，进行预处理，选择其中剧情简介和短评齐全的 1634 个电影数据，经过次序打乱后重新组织进行本次实验。整理后训练集和测试集样本数量如表 3 所示：

实验使用复旦大学的 NLP 工具包 FudanNLP（Version 1.5）进行分词处理。分类算法是

[1] http://code.google.com/p/fudannlp/

支持向量机 SVM，使用 Joachims 的 SVM-light 工具包中的 SVM 分类器进行分类训练。使用简介与短评的直接融合，投票规则，加法规则分别进行分类器的融合，并将融合结果与单一使用简介、短评和关键词进行分类的结果进行对比，采用准确率（precision）、召回率（recall）和 F 值（F-measure）这三个指标对模型推荐结果进行评价。

表 3　训练集和测试集数量

Table 3　Number of training set and test set

标签	训练集		测试集	
	包含标签	不包含标签	包含标签	不包含标签
美国	375	375	360	360
喜剧	239	239	240	240
爱情	225	225	225	225
剧情	186	186	185	185
动画	184	184	185	185
科幻	188	188	180	180
经典	162	162	160	160
动作	138	138	135	135
搞笑	106	106	100	100
香港	79	79	80	80

5．2 实验结果与分析

实验结果如表 4 和图 2 所示，Summary 表示使用剧情简介进行标签推荐，Comments 表示使用短评信息进行标签推荐，Keyword 表示使用关键词搜索匹配的方法进行标签推荐，Sum+Com 表示使用剧情简介和短评信息的直接融合进行标签推荐，Vote 表示使用投票规则融合分类器进行标签推荐，Summation 表示使用加法规则融合分类器进行标签推荐。

表 4　各分类器及融合分类器的准确率、召回率、F1 值比较

Table 4　Comparison of the accuracy, recall and F1 of the classifier and classifier combination

Classifier	P	R	F1
Summary	77.45%	74.37%	75.76%
Comments	71.27%	70.77%	70.06%
Keyword	87.32%	34.14%	48.32%
Sum+Com	80.22%	80.00%	79.75%
Vote	80.40%	79.84%	79.76%
Summation	80.67%	79.04%	79.46%

从表 4 的结果可以看出，使用直接融合、投票规则、加法规则的方法都比单一使用 Summary 或 Comments 的分类方法有较大的提高，F1 平均值比 Summary 分别提高了 3.99%，4.00%，3.70%，比 Comments 分别提高了 9.69%，9.70%，9.40%，比 Keyword 分别提高了 31.43%，31.44%，31.14%。从而证明融合两方面信息比只用一方面信息更有效。

从图 2 的结果可以看出，不同的标签分类结果有一定的差异，表示属性的名词类标签，如“美国”、“爱情”、“动画”、“科幻”、“动作”、“香港”在 Summary 中大多已被描述，分类结果良好。而形容词性的标签，如“经典”、“搞笑”，在 Summary 中一般不包含，在短评中却常被用户描述，通过 Summary 与 Comments 的融合也获得了比较好的结果。而名词“剧情”既不表示电影的属性，也不是对电影的评价，而是电影的二级属性，在 Summary 和 Comments 中被描述的可能性都较低，分类结果低于平均水平。

实验证明，融合方法对单一的训练具有更好的性能，可以充分利用 Summary 和 Comments 的优势，对电影标签进行有效的自动推荐。

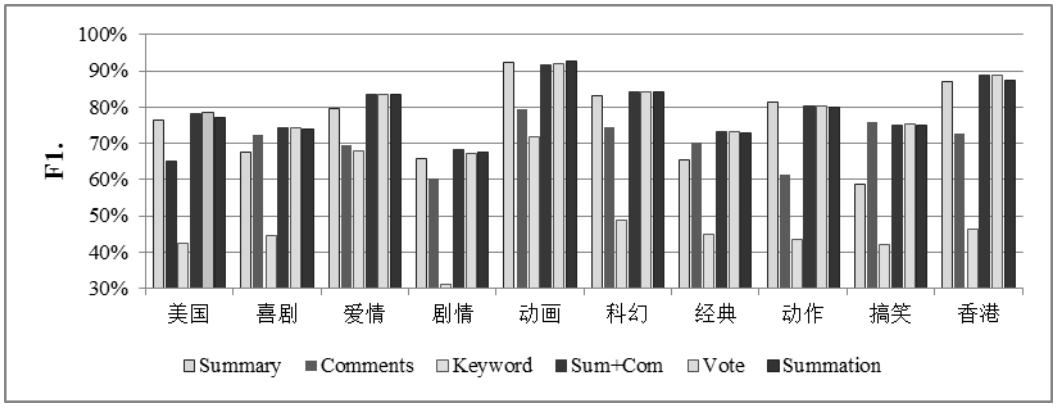

图 2　各分类器及融合分类器在不同标签上的 F1 值柱状比较

Fig. 2 Comparison of different ensemble learning approach with different tags

6　结论和下一步工作

本文提出利用剧情简介和短评信息各自的优势，使用剧情简介和短评的信息进行电影标签自动推荐。针对剧情简介和短评的基分类结果，分别使用直接融合、投票规则、加法规则等方法进行分类器融合，实验结果表明融合后的结果都明显的好于单一信息进行标签推荐的方式。下一步计划进一步探讨其他的融合方法的使用，以及利用电影相关的其他信息，如影评、问答、相关电影、相似电影等信息进行标签推荐的研究。

参考文献

[1] Eck D, Lamere P, Bertin-Mahieux T, and Green S. Automatic Generation of Social Tags for Music Recommendation[C]//NIPS-2007, 8: 385-392.

[2] Yanbe Y, Jatowt A, Nakamura S, and Tanaka K. Can Social Bookmarking Enhance Search in the Web?[C]// Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, 2007: 107-116.

[3] Zhou T C, Ma H, Lyu M, and King I. UserRec: A User Recommendation Framework in Social Tagging Systems[C]//AAAI-2010: 1486-1491.

[4] Hotho A, Jäschke R, Schmitz C, and Stumme G. Trend detection in folksonomies[M]// Semantic Multimedia. Springer Berlin Heidelberg, 2006: 56-70.

[5] Wetzker R, Zimmermann C, Bauckhage C, and Albayrak S. I tag, you tag: translating tags for advanced user models[C]// WSDM-2010: 71-80.

[6] Mirizzi R, Ragone A, Di Noia T, and Di Sciascio E. Semantic tags generation and retrieval for online advertising[C]// CIKM-2010: 1089-1098.

[7] Ohkura T, Kiyota Y, and Nakagawa H. Browsing System for Weblog Articles based on Automated Folksonomy[C]// WWW-2006:25-27.

[8] Mishne G. AutoTag: a collaborative approach to automated tag assignment for weblog posts[C]// WWW-2006:953-954.

[9] Blei D M, Ng A Y, and Jordan M I. Latent Dirichlet Allocation[J]// Journal of Machine Learning Research,2003:993-1022.

[10] Hofmann T. Probabilistic Latent Semantic Indexing[C]// SIGIR-1999:50-57.

[11] Blei D, and McAuliffe J. Supervised topic models[C]// NIPS-2008,20:121-128.

[12] Si X, and Sun M. Tag-LDA for Scalable Real-time Tag Recommendation[J]//Journal of Computational Information Systems,2009:6(2).

[13] Herlocker J L, Konstan J A, Borchers A, Riedl J. An algorithmic framework for performing collaborative filtering[C]// SIGIR-1999: 230-237.

[14] Herlocker J L, Konstan J A, Terveen L G, and Riedl J. Evaluating collaborative filtering recommender systems[C]// ACM Transactions on Information Systems (TOIS), 2004, 22(1): 5-53.

[15] Resnick P, and Varian H R. Recommender systems[C]// Communications of the ACM, 1997, 40(3):

56-58.

[16] Nakamoto R, Nakajima S, Miyazaki J, and Uemura S. Tag-Based Contextual Collaborative Filtering[J]// IAENG International Journal of Computer Science, 2007, 34(2).

[17] Niwa S, and Honiden S. Web Page Recommender System based on Folksonomy Mining [C]// Information Technology: New Generations-2006: 388-393.

[18] Gemmell J, Shepitsen A, Mobasher B, and Burke R. Personalizing navigation in folksonomies using hierarchical tag clustering[M]// Springer Berlin Heidelberg, 2008: 196-205.

[19] Santos-Neto E, Ripeanu M, and Iamnitchi A. Tracking user attention in collaborative tagging communities[J]. International ACM/IEEE Workshop on Contextualized Attention Metadata: Personalized Access to Digital Resources, 2007.

[20] Liu X, Wang Y, Liu Z, and Xie M. Tag recommendation based on continuous conditional random fields[C]// Information Management, Innovation Management and Industrial Engineering, 2009 International Conference on IEEE, 2009, 3: 475-480.

[21] Jäschke R, Marinho L, Hotho A, and Schmidt-Thieme L. Tag recommendations in social bookmarking systems[J]. Ai Communications, 2008, 21(4): 231-247.

[22] Rendle S, Balby Marinho L, Nanopoulos A, and Schmidt-Thieme L. Learning optimal ranking with tensor factorization for tag recommendation[C]// KDD- 2009: 727-736.

[23] Kittler J, Hatef M, Duin R P W, and Matas J. On combining classifiers[J].// IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(3): 226-239

作者简介：

 褚晓敏（1981——），女，博士研究生，主要研究领域为自然语言处理。Email:xiaomin.chu@gmail.com；

 王中卿（1987——），男，博士研究生，主要研究领域为自然语言处理。Email:wangzq.antony@gmail.com；

 朱巧明（1963——），男，博士，教授，主要研究领域为自然语言处理。Email:qmzhu@suda.edu.cn；

周国栋（1967——），男，博士，教授，主要研究领域为自然语言处理。Email:gdzhou@suda.edu.cn。