# 硕士学位论文

# 面向微博用户的标签自动生成技术研究

# RESEARCH ON AUTOMATIC GENERATION OF TAGS FOR MICROBLOG USERS

谢毓彬

哈尔滨工业大学

2012 年 6 月

国内图书分类号：TP391.2　　学校代码：10213
国际图书分类号：681.37　　密级：公开

工学硕士学位论文

# 面向微博用户的标签自动生成技术研究

硕 士 研 究 生：谢毓彬
导　　师：刘挺教授
申 请 学 位：工学硕士
学　　科：计算机科学与技术
所 在 单 位：计算机科学与技术学院
答 辩 日 期：2012 年 6 月
授予学位单位：哈尔滨工业大学

Classified Index: TP391.2
U.D.C: 681.37

Dissertation for the Master Degree in Engineering

# RESEARCH ON AUTOMATIC GENERATION OF TAGS FOR MICROBLOG USERS

**Candidate:** Xie Yubin
**Supervisor:** Prof.Liu Ting
**Academic Degree Applied for:** Master of Engineering
**Speciality:** Computer Science and Technology
**Affiliation:** School of Computer Science and Technology
**Date of Defence:** June, 2012
**Degree-Conferring-Institution:** Harbin Institute of Technology

# 摘　要

近年来，微博服务作为新型的互联网应用，受到了越来越多用户的关注。在自然语言处理、信息检索和社会计算等相关领域，针对微博的研究工作也在逐渐开展和积累中。微博用户标签，作为描述用户兴趣爱好、职业领域特征等的载体，在用户组织和搜索，挖掘用户兴趣、实现微博上的个性化等方面有着重要的作用。

本文着眼于基于微博内容的用户标签自动生成，借助对内容的分析，生成能够体现用户兴趣的标签。

本文通过新浪微博 API 随机获取了百万级规模的标签相关数据，用于分析用户标签在统计、语义等方面的特征。同时，我们对基于文本的标签源：用户的原创、转发、评论和收藏微博的语义相似度及其对反映用户兴趣的贡献进行了实验和分析，结果表明标签源间的语义相似度并不高；而转发微博更能体现用户兴趣，评论最差，从而也确定了本文中生成标签的文本来源。

本文从生成标签的不同粒度出发，分别从基于关键词和基于类别的角度自动生成微博用户标签。对生成结果的评价准则有两条：一是生成结果是否准确体现了用户兴趣；二是生成结果是否适合作为用户标签。

在基于关键词的生成方法中，引入了基于 TextRank 的标签生成方法，通过分析微博中词语的共现关系，构建词语网络，抽取较为重要的词用于标签生成。为了使生成的标签能在更多的维度上体现用户兴趣，接着提出了基于聚类分析的生成方法，从较重要的聚类簇中提取代表词用于标签生成。实验表明，两种方法都优于我们的 baseline。同时，我们也对两种方法进行了讨论、对比和分析。

在基于类别的生成方法中，将用户感兴趣的若干个类别作为其标签。首先提出了基于短文本分类的标签生成方法，人工构建目标分类体系及微博训练语料，识别出用户感兴趣的类别作为标签。随后，我们在更细的粒度上为用户打标签：利用百度百科具有三层分类信息的词条资源，识别出用户关注的类别作为标签。实验表明，两种方法生成标签的准确率均能达到 70%左右。同时，我们也对这两种方法进行了讨论、对比和分析。

**关键词**：微博用户标签；TextRank；聚类分析；文本分类；百度百科

# Abstract

In recent years, micro-blogging services have attracted more and more attention as a new type of web applications. Research focused on microblog is gradually carried out and in accumulation in related fileds such as natural language processing, information retrival and social computing. Tags for a microblog user, as the description for his/her interests, concerns and occupational characteristics, are playing an important role in user indexing and searching, personalization and so on in microblog.

This work focuses on user tagging which is based on micrblog content. User tags are automatically generated with the analysis of microblog content.

Millions of user tags related data are crawled via sina microblog API, which are then used to analyze the characteristics of user tag in statistics and semantic respects. Meanwhile, experiments and analysis have been done about the semantic similarity and the contribution to reflect user interests of tag sources based on text, the content user writes, repost, comment or favorites. The results show that microblog content users repost reflect user interests most, and comment least. And content used to generate user tags are the decided based on those experimental results.

Proposed keywords-based and category-based approaches based on the view of different granularity of user tag. There are two criteria to evaluate generation results. Generations should accurately reflect user interests and be suitable as user tags.

Generation based on TextRank is introduced in keywords-based approach. Important words are extacted for user tagging by analyzing the co-occurrence of words in microblog and then constructing word network for them. Generation based on cluster analysis is then proposed to discover user interests on more dimensions, which mainly extract representations from clusters for user tagging. The experimental results show two keywords-based approaches pefrom better than baseline approach. And comparisions and discussions are made between these two approaches.

Categories which users concerned are used as user tags in category-based approaches. Generation based on short text classification is proposed. Target classes and microblog corpus are construced to recognize user interests. Baidu Encyclopedia based approach is then introduced to for user tagging with its three-level-categroy information of word items. The experimental results show the accuracy of user tagging of these two category-based approaches achieves 70

percent on test data. Meanwhile, comparisions and discussions are made between these two approaches.

**Keywords:** Tags for Microblog Users, TextRank, Cluster Analysis, Text Classification, Baidu Encyclopeida

# 目录

摘　要 ..... I
Abstract ..... II
第 1 章 绪　论 ..... 1
1.1 课题背景 ..... 1
1.2 课题研究目的和意义 ..... 2
1.3 国内外相关研究现状与分析 ..... 4
1.3.1 社会化标签 ..... 4
1.3.2 微博用户标签 ..... 6
1.4 本文的主要研究内容及章节安排 ..... 8
第 2 章 标签相关数据的获取及分析 ..... 9
2.1 引言 ..... 9
2.2 标签相关数据的获取 ..... 9
2.2.1 微博 API 介绍 ..... 9
2.2.2 基于微博 API 的数据获取 ..... 12
2.3 标签相关数据分析 ..... 12
2.3.1 数据集 ..... 12
2.3.2 标签的若干特征 ..... 12
2.4 本章小结 ..... 16
第 3 章 基于文本的标签源分析 ..... 18
3.1 引言 ..... 18
3.2 用户产生标签源的行为及统计性质 ..... 18
3.3 标签源的语义相似度 ..... 19
3.3.1 词语级别的相似度 ..... 19
3.3.2 主题级别的相似度 ..... 21
3.4 标签源选择对反映用户兴趣的影响 ..... 22
3.4.1 方案 ..... 22
3.4.2 数据 ..... 23
3.4.3 结果 ..... 23
3.5 本章小结 ..... 27
第 4 章 基于关键词的标签自动生成 ..... 28
4.1 引言 ..... 28
4.2 基于 TextRank 的生成方法 ..... 28
4.2.1 TextRank 概述 ..... 28
4.2.2 生成方法 ..... 29
4.2.3 实验与结果分析 ..... 31
4.3 基于聚类分析的生成方法 ..... 34
4.3.1 关键技术与原理 ..... 34
4.3.2 生成方法 ..... 36
4.3.3 实验与结果分析 ..... 38

4.4 两种方法对比与分析 ..... 40
4.5 本章小结 ..... 42
第 5 章 基于类别的标签自动生成 ..... 43
5.1 引言 ..... 43
5.2 基于短文本分类的生成方法 ..... 43
5.2.1 关键技术与原理 ..... 43
5.2.2 生成方法 ..... 47
5.2.3 实验与结果分析 ..... 49
5.3 基于百度百科的生成方法 ..... 55
5.3.1 百度百科介绍 ..... 55
5.3.2 生成方法 ..... 56
5.3.3 实验与结果分析 ..... 58
5.4 两种方法对比与分析 ..... 60
5.5 本章小结 ..... 60
结 论 ..... 61
参考文献 ..... 63
攻读学位期间发表的学术论文 ..... 67
哈尔滨工业大学学位论文原创性声明及使用授权说明 ..... 68
致 谢 ..... 69

# 第1章 绪 论

## 1.1 课题背景

近几年来，微博（即微博客），作为新兴的互联网应用，受到了越来越多人的欢迎，无论是各领域中的名人还是普通的网民，都可能拥有微博账户。

自 2006 年成立以来，截至 2010 年 1 月份，来自美国的最早、最著名的微博 Twitter[1]在全球已拥有 7500 万注册账户。据 Twitter 账户独立跟踪机构 Twochars 发布的研究报告显示，其注册账户预计将在今年年底前达到 9 亿[1]。

在我国，2009 年以前，具有微博色彩的产品例如饭否、腾讯滔滔便已经开始出现。2009 年 8 月份，中国最大微博服务新浪微博横空出世。2010 年，国产微博迎来了春天，中国四大门户网站——新浪、腾讯、搜狐、网易均开设了自己的微博客平台[2]。2012 年 1 月，我国互联网络信息中心（CNNIC）在京发布了《第 29 次中国互联网络发展状况统计报告》。《报告》数据显示，截至 2011 年 12 月底，我国微博的用户数目已经达到 2.5 亿，相比上一年增幅高达 296.0%，网民使用率为 48.7%。微博作为用户增长最快的互联网应用模式，已成为近一半中国网民使用的重要互联网应用。

随着微博的迅猛发展，与之相关的课题研究也逐渐成为了学术界的热点。为了能够在本文中，更好地阐述我们的课题，我们首先将对与课题相关的一些微博特征、功能等进行简单的介绍：

在微博平台中，一般来说，用户可以通过与微博账号连接的客户端，随时随地发布长度不超过140个字的简短内容，也称为微博。

与传统的社交网络明显不同的是，微博的关注机制可单向也可双向。也就是说，关注是一种单向的，而且不需要得到对方确认的关系。倘若用户 A 对另一个微博用户 B 感兴趣，加关注后，用户 A 就成为了用户 B 的粉丝(follower)，并且可以实时地浏览到被关注者（followee）B 发布的任何一条微博。

用户登陆微博平台后，可以实时地浏览到他/她关注的人发布的内容，这样的微博列表通常被称为 Timeline。当用户对某条微博感兴趣时，可以进行评论、转发或者收藏。

微博用户被允许使用自造的、长度不限的词语或者短语来描述、标识自己。这就是所谓的标签功能。图 1-1 中的两个图，分别是国内新浪和腾讯两大微博的个人标签设置页面。它们对个人标签的个数都提出了限制，每个人至多为自己打上 10 个标签。我们可以看到，这两个页面，向用户表述了标签的功

[1] http://www.twitter.com

能和意义，同时均附有标签推荐功能。

在本文中，我们将专注于微博用户的个人标签，研究其自动生成方法。


(a) 新浪微博标签设置页面


(b) 腾讯微博标签设置页面

图 1-1 微博的标签设置页面

## 1.2 课题研究目的和意义

结合微博平台对用户标签的表述，本文中，我们不妨把用户标签定义为能够表露用户喜欢、关心的内容例如兴趣爱好、职业领域特征的较简短的文本。微博用户标签自动生成，便是从可利用的资源中自动生成短文本例如词语、短语来描述用户关心的内容、兴趣点等。

一般来说，我们关注某个用户，常常是因为和自身有着一定的共同点，对对方的某些方面感兴趣，例如都从事自然语言处理研究或者都是羽毛球的爱好者等等，而且，我们往往都能从对方发布的微博中浏览到自己对应感兴趣的内容。相应地，一旦某个用户将自然语言处理、羽毛球设置成了个人标签，其他拥有相同兴趣爱好的人就可以通过标签搜索进而关注他。


(a) 新浪微博标签搜索


(b) 腾讯微博标签搜索

图 1-2 微博的标签搜索页面

标签相当于一张个人名片，通过它，我们可以搜索到一大批具有相同兴趣爱好、职业领域的人，丰富自己信息获取的渠道。此外，一旦我们为自己打上了标签，也能更方便地被其他人搜索到，提高自身影响力，扩充自己在微博平台上的社交网络。图 1-2 分别是新浪、腾讯提供的标签搜索功能。

微博文本短小（一般限制在 140 个字以内），与手机、即时信息服务软件的连接使得微博发布的门槛很低，用户可以随时随地进行更新。通过关注某一个用户，无论对方是你真实生活中的朋友，还是专业领域内的精英，甚至是媒体官方微博……我们可以实时地获取他们发布的最新信息。这些信息可能包括日常感悟、心情记录、产业资讯、新闻等等。随着我们关注的人数越多，这股信息流将越大。因此，我们将面临“信息过载”的问题。即我们看到的微博多数可能并非自己真正关心的、感兴趣的。简单来说，就是信息太多，一时看不过来。

针对信息过载的问题，人们开始逐渐关注微博上的个性化。个性化的核心往往在于为用户建立较为准确的兴趣模型。微博用户标签，作为用户自定义的描述，带有明显的用户兴趣信息，为个性化提供了重要的作用。微博用户标签自动生成，课题的关键点就在于识别、描述微博用户的兴趣。该课题的研究亦有助于微博用户兴趣建模、微博个性化推荐等研究。

通过标签，微博平台服务者可以更好地了解用户的兴趣和视角，以便让服务更加个性化，进而为用户提供更加精准的推荐和广告服务，为自身带来盈利。而用户，也能获得更佳的用户体验。

因此，本课题是具有研究和应用前景的。我们试图识别出微博用户的兴趣，并对其进行描述，自动生成个人标签。

## 1.3 国内外相关研究现状与分析

微博作为一种新兴的媒体，相关的研究在近几年才逐渐成为学术界的热点，单独就微博用户标签自动生成开展的课题研究更是很少。因此，本文将介绍与其相关的一些研究工作。

### 1.3.1 社会化标签

微博用户标签，可能会让很多人联想到另一个词语：社会化标签。我们将简单介绍这两个听起来有关联且相似的领域，并将通过介绍，简单阐述这两者及其两者的自动生成之间的异同。

社会化标签（social tagging）因为诸如 Delicious 和 Flickr 之类的网站的出现而变得流行。例如，Delicious 允许用户对关注的 URL 进行描述、标注；Flickr 是一家主要提供图片服务的网站，它允许用户对自己或者他人上传的图片进行标注；在博客类网站如 Wordpress，Livejournal 以及国内的新浪、搜狐

等博客，博主可以为自己发表的博文添加标签；在社交型网站如 Facebook，用户不仅可以对图片进行标注，还能为其他信息打上“Like”之类的标签。

上述些网站有着一个共同的特点，就是允许用户自由地使用自定义的或者网站提供的词语、短语等作为标签对网络资源进行标注，从而也起到了组织、分类、分享互联网资源的作用。标签是用户根据自身喜好对关注的特定资源进行的描述、注释，能够帮助用户对资源进行更便捷地浏览、检索、组织、管理，以及与其他网民分享。当然，标签的添加在一定程度上也是耗费精力的事情，因此，很多的社会标注系统也提供标签推荐服务，对特定的资源提供一系列候选的标签，而相对应地，在学术界也有社会化标签推荐相关的研究。

我们来简单地阐述下社会化标签的相关定义。一般，我们将一个特定用户在某个社会标注系统中的所有标注称为他的 personomy，而所有用户的 personomy 就被称为 folksonomy[3]。Thomas Vander Wal 在 2004 年第一次提出了 folksonomy 一词，用以描述用户群体自发地为特定的资源进行标注。他提出，标签不仅仅是对实体的描述，也是分类的过程[4]。在国内，我们也常称它为大众分类、民俗分类、社会化分类、自由分类等。

社会化标签一般拥有以下特点[5]：

1）自由性和共享性。一般来说，网络用户可以对自己感兴趣的任意的互联网资源进行标注，添加一个或者多个标签，这种行为是自由的、自主的。而其他用户可以浏览他人的标签，并进行修改、添加。当然，folksonomy 的定义者曾在文献[6]中指出，社会标注有广义和狭义之分。狭义的社会标注，标签的标注、修改行为需要得到一定的允许。

2）动态更新。随着用户群体的协同标注的积累，资源的标签的信息也在不断更新，丰富着整个社会化标签的集合。

一个社会化标签集合通常由（用户、标签、资源）三元组来描述[7]。而关于社会化标签的研究也常常围绕着这三者，或者这三者之间的关系进行。

在文献[8]中，作者对近年来的社会化标签技术进行了细致而全面的总结分析。Folksonomy 涉及的研究课题包括用户标注行为、标签本身、标签的自动生成或者推荐、标签的可视化等等方面。

自从 2005 年以来，已经有很多工作专注于分析人们的标注动机以及标签的类型。例如，人们通过标签为以后的资源检索提供便利、通过标签表达自己的观点和情感，热门的标签还能够吸引他人的注意。标签的类型可分为基于内容的标签（描述资源的实质内容）、基于上下文的标签（描述关于该资源的地点、时间等）、属性标签（例如 scray、funny、stupid）等等[9-11]。

为了更好地理解社会化标签数据，很多工作分析了数据的分布、语义等各种方面的性质。有人将 Flickr 的标签映射到 WordNet，仅使用简单的字符串匹配[12]。他们发现 51.8%的标签可以通过这种简单的映射被分类。有人试图将大

量的标签转化可导航的层次化的结构，从而进行层级化的表示[13]。

当然，还有很多的研究工作是针对社会化标签自动生成模型或者推荐算法的。文献[14-16]将标签推荐的方法大体分为两类，一是基于图的方法或者说是协同推荐方法，二是基于内容的方法。Jaschke[17]等提出了类似 PageRank 的方法 FolkRank 进行标签推荐。当然，更多的是基于 K 近邻、矩阵分解的算法。协同推荐的算法往往会面临“冷启动”的问题，即一个新用户加入系统后，系统往往难以对其进行推荐。基于内容的方法，正好对此有了补充。概率模型、语言模型往往被用于标签的自动生成。

由上述介绍，相信我们已经可以体会到微博用户标签与社会化标签之间的差异。在微博用户标签中，（用户、资源、标签）三元组，实际上资源也就是用户本身。从某种角度来说，微博用户标签应该算是社会标签的一种特殊情况。在国内微博这个平台上，用户只允许为自己添加至多 10 个的标签，无权为其他用户添加标签。因此，关于微博用户标签自动生成的研究，我们缺乏如社会化标签那样的用户-资源-标签的三部图，也缺乏用户的标注历史，作为我们自动生成标签的资源，可利用的资源可谓比较匮乏。

尽管微博用户标签和社会化标签之间存在很大的差异，但是，我们仍然认为社会标签中关于用户标注动机、标签属性如分布和语义、标签的自动生成和推荐方面的研究工作对于我们的课题研究具有借鉴价值。

### 1.3.2 微博用户标签

自 twitter 诞生以来，微博在吸引大批用户的同时也吸引了越来越多的学者。最初，人们还只是综合地研究微博的一些特征，例如用户的性别、年龄、地理分布，用户关注、粉丝、发布微博的数量等等；随后，针对微博各方面的较为深入的研究也逐渐增多。其主要研究方向包括事件和话题的发现、跟踪、传播和预测，微博社交网络挖掘，情感倾向性分析，用户影响力分析，微博用户建模和个性化分析等等[18-30]。

本文研究的微博用户标签自动生成，与微博的内容分析、建模，微博用户的兴趣分析等有着较为紧密的联系，因此，我们下面将介绍与这些方面有关的现有的研究工作。

赵鑫等人[31]利用主题模型对 Twitter 与以 New York Times 为例的传统媒体的内容做出了比较。他们对标准的 LDA 模型做出了改进，提出了 Twitter-LDA，对每一条简短的 tweet 只赋予一个主题。实验对比发现，Twitter-LDA 建立的模型的主题、含义更加清晰。通过大量的数据分析，他们发现在 Twitter 上，人们倾向于谈论与家庭、生活相关的话题，而在新闻类型的话题上体现出的兴趣相对地更低，同时，Twitter 的内容涵盖了更多的实体类型的

主题。

Liangjie Hong 等[32]深入研究了在微博环境如何使用数据集训练标准的主题模型，以及模型的质量、有效性如何。文中，作者使用三种不同的策略来训练模型：将单条微博当成一篇文档的 MSG 策略、将同一作者的所有微博聚合起来的 USER 策略，将涉及相同 hashtag 的微博聚合在一起的 TERM 策略。实验表明，MSG 和 TERM 策略训练的模型具有更高的主题分布相似度，然而 USER 策略在微博分类等场景中具有更佳的表现。

Daniel Ramage 等[33]利用 Labeled-LDA 对 Twitter 的内容和用户建模，并用于微博排序、用户推荐等任务，表现出了不错的性能。

来自荷兰 DELFT 理工大学的 Fabian Abel 等人在用户建模上做了很多研究工作。他们通过提取微博中的 hashtag、实体等，并与当前主流媒体如 CNN、CBC、New York Times 相链接，丰富微博的语义[34]。据 Kwak 等人[18]的研究表明，在 twitter 超过 85%的微博是与新闻相联系的。而在文献[34]中，他们的实验结果表明，约 15%的微博文本可以通过实体与新闻文章建立关系，而且，给定任意一个新闻文章中提及的实体，找到与其相关的微博的准确率超过 75%。同时，他们发现建立 tweet-news 的关系，能够对用户建模起到重大的作用。文献[35]中，他们进一步利用传统媒体中的新闻、微博中的 hashtag 等，提出了基于 Twitter 的用户建模框架，并应用于推荐任务中。同时，他们的研究工作还引入了时间的因素，通过为用户构造长期和特定时间段的模型，观察用户兴趣的变化。文献[36]中，基于他们的前期研究工作，提出了自己创建的基于 Twitter 的用户模型应用 TUMS。给定一个 Twiiter 用户，收集该用户发布的所有微博，丰富语义，返回用户建模结果，并对其可视化。

Matthew Michelson 和 Yegin Genc 等人将微博内容与维基百科资源结合进行研究。文献[37]将用户发布的微博中提及的实体经过消歧等处理后映射到维基百科的某个类别节点上，经过投票策略可得到用户最感兴趣的维基百科类别节点。文献[38]目的在于对单条微博进行分类。作者同样提取微博中的实体，得到每个实体对应的维基百科类别节点。因为节点是具有层次的，作者通过一个基于路径的算法得到每条微博最终的类别。

此外，还有少数研究工作是直接针对微博用户标签的。

Wei Wu 等人曾利用基于 TextRank 的方法抽取用户微博的关键词作为用户标签[39]。Theodoros Lappas 等人[40]利用社会支持网络（Social Endorsement Networks）来挖掘 Twitter 用户的标签。作者认为，不同用户可能因为对某个特定的用户的不同方面感兴趣而关注他，因此可以通过分析该用户的粉丝网络来获取标签。文中抽取用户的粉丝发布的微博建立改进的主题模型，挖掘标签。但是，该方法主要针对的是粉丝人数较多的微博用户，实验是在粉丝数超过 2000 的用户中进行。Yuto Yamaguchi 等[41]利用 Twiiter 用户的分组名称来给

用户添加标签。实验取得了不错的性能，但是正如作者所说，挖掘的标签存在大量的同义标签。后续需要进一步地处理。

## 1.4 本文的主要研究内容及章节安排

本文主要针对中文微博，进行用户标签自动生成技术的研究。本文从微博内容分析的角度出发，来展开研究，旨在生成能够体现用户兴趣的标签

本文的章节安排如下：

第一章，首先，阐明本研究工作的背景、目的和意义。由于与微博用户标签自动生成直接相关的研究工作还较少，因此，接下来详细介绍了与本课题较为相关的国内外研究现状，包括社会化标签、微博内容分析等等。最后，介绍整篇论文的主要研究内容及各个章节的安排。

第二章，主要介绍与微博用户标签相关的数据获取与分析工作。首先，介绍数据的获取方式。接下来，通过获取的数据，我们进行统计、分析，挖掘出用户使用标签功能的规律，标签数据的若干统计、语义特征。

第三章，由于本文着眼于基于内容分析的标签生成，因此，我们还借助真实的数据和实验，探索了用户的原创、转发、评论和收藏行为和相应产生的内容，及其对反映用户兴趣的贡献，确定后续工作使用的文本标签源。

第四章，从关键词的角度为用户生成标签。本章分别介绍了基于 TextRank 和基于聚类分析的两种生成方法，并就两种方法开展了实验，做了对比与分析。

第五章，从类别的角度为用户生成标签。首先，介绍了基于短文本分类的方法，从单层次上为用户生成标签。接下来，介绍了基于百度百科的方法，借助丰富的词条信息，为用户生成多层次类别的标签。我们就两种方法进行了实验，并对结果进行了分析。

# 第2章 标签相关数据的获取及分析

## 2.1 引言

面向微博用户的标签自动生成技术研究，研究的对象是微博用户标签。对标签相关的数据进行较为全面的分析，有利于我们研究课题的深入开展。

虽然说早在 5 年之前，中国就有一家带有微博色彩的网站饭否开张，但是 2010 年随着新浪、腾讯、网易、搜狐四大门户网站微博的开设，中文微博可以说才迎来了自己的春天，进入大众的视野。针对中文微博的相关研究也才起步不久。因此，我们也有必要深入了解下我们需要研究的对象。

新浪微博目前可以说是中国最受欢迎的微博服务，相关数据表明，新浪微博的注册用户已经超过 3 亿。本文也将借助新浪微博相关的数据，展开研究。

## 2.2 标签相关数据的获取

国内微博与 twitter 一样，为广大用户提供了 API，方便获取用户个人信息、发布的微博文本、社交关系等等，用于创建有趣的微博应用或者其他无害的用途。下面，我们将以新浪微博 API 为例，展开介绍。

### 2.2.1 微博 API 介绍

新浪的微博接口按照 IO 操作来分类，分为读取和写入接口两类；按照涉及的信息来分，可分为微博接口、评论接口、用户接口、标签接口等十二大类，具体如图 2-1 所示。我们通过 API 请求得到的信息，通常以 XML 或者 JSON 的格式返回，其中包含各种重要的字段。下面，将分别介绍本文涉及的几个主要的接口。


图 2-1 新浪微博接口类型

**（1） 用户信息接口**

通过该接口请求返回的信息内容包含了诸多标识、描述用户特征的信息。用户 ID 字段，每一个微博用户都将以一串唯一的、与他人不同的数字来代表；screen_name 字段，是用户在微博上显示的昵称；gender 字段，用于填写用户的性别；而用户的创建时间、关注数、粉丝数、微博数、收藏数分别用 created_at、friends_count、followers_count、statuses_count、favourites_count

字段来表示。表 2-1 是详细的用户信息接口字段说明。

表 2-1 用户信息接口字段说明

| 字段 | 字段说明 |
|---|---|
| id | 用户 UID |
| screen_name | 用户昵称 |
| name | 友好显示名称 |
| province | 用户所在地区 ID |
| city | 用户所在城市 ID |
| location | 用户所在地 |
| description | 用户描述 |
| url | 用户博客地址 |
| profile_image_url | 用户头像地址 |
| domain | 用户的个性化域名 |
| gender | 性别，m：男、f：女、n：未知 |
| followers_count | 粉丝数 |
| friends_count | 关注数 |
| statuses_count | 微博数 |
| favourites_count | 收藏数 |
| created_at | 创建时间 |
| following | 当前登录用户是否已关注该用户 |
| allow_all_act_msg | 是否允许所有人给我发私信 |
| geo_enabled | 是否允许带有地理信息 |
| verified | 是否是微博认证用户，即带 V 用户 |
| allow_all_comment | 是否允许所有人对我的微博进行评论 |
| avatar_large | 用户大头像地址 |
| verified_reason | 认证原因 |
| follow_me | 该用户是否关注当前登录用户 |
| online_status | 用户的在线状态，0：不在线、1：在线 |
| bi_followers_count | 用户的互粉数 |
| status | 用户的最近一条微博信息字段 |

（2） **微博信息接口**

我们主要是通过该接口请求用户发布的微博内容，用于稍后的标签自动生成。返回的信息中，包括唯一标识该条微博的 ID 字段，该条微博携带的真正的文本信息 text 字段，该条微博的评论数、转发数等等。当该条微博是转发的微博时，字段中还将嵌套被转发的微博的信息，用 retweeted_status 字段来表示。表 2-2 是详细的微博信息接口字段说明。

（3） **用户标签接口**

通过该接口，我们可以得到指定用户的标签列表。一个标签仅由标识它的 id 以及值 value 即文本信息字段来表示。图 2-2 是从接口中获取的某个用户的标签列表 XML 格式示例。

表 2-2 微博接口字段说明

| 字段 | 字段说明 |
|---|---|
| Idstr | 字符串型的微博 ID |
| created_at | 创建时间 |
| Id | 微博 ID |
| Text | 微博信息内容 |
| Source | 微博来源 |
| Favorite | 是否已收藏 |
| Truncated | 是否被截断 |
| in_reply_to_status_id | 回复 ID |
| in_reply_to_user_id | 回复人 UID |
| in_reply_to_screen_name | 回复人昵称 |
| Mid | 微博 MID |
| bmiddle_pic | 中等尺寸图片地址 |
| original_pic | 原始图片地址 |
| thumbnail_pic | 缩略图片地址 |
| reposts_count | 转发数 |
| comments_count | 评论数 |
| Annotations | 微博附加注释信息 |
| Geo | 地理信息字段 |
| retweeted_status | 源微博即被转发的微博信息 |
| User | 微博作者的用户信息字段 |

```
<?xml version="1.0" encoding="UTF-8"?>
<tags>
  <tag>
    <id>547035</id>
    <value>网络交易</value>
  </tag>
  <tag>
    <id>51580</id>
    <value>C2C</value>
  </tag>
  <tag>
    <id>11547</id>
    <value>外星人</value>
  </tag>
  <tag>
    <id>787292</id>
    <value>智能设备</value>
  </tag>
  <tag>
    <id>1358</id>
    <value>电子商务</value>
  </tag>
  <tag>
    <id>104715</id>
    <value>艺术品交易</value>
  </tag>
  <tag>
    <id>2243</id>
    <value>互联网产品</value>
  </tag>
</tags>
```

图 2-2 用户标签列表示例

（4）社交关系接口

通过该接口，我们可以得到某个指定用户的关注和粉丝列表。关注和粉丝依然用唯一标识他们的 ID 字段来代表。

### 2.2.2 基于微博 API 的数据获取

获取数据的方式有两种，一是通过申请开发微博应用的方式，使用相应 API 获取数据；二是通过普通网页爬虫的方式，使用网页解析手段获取相关数据，但是由于隐私的设置，某些信息难以爬取。因此，本文采用第一种方式，基于微博 API 来获取数据。

国内微博开放平台对数据获取的权限作出了严格的限制。以新浪微博为例，对于一个应用的一个用户而言，每小时仅能调用 150 次 API；对于一个 IP 地址来说，每小时仅能调用 10000 次 API。另外，我们每一次 API 请求，获取的信息量也是极其有限的。例如，我们通过 statuses/user_timeline 这个地址来抽取用户曾经发布的微博列表，每一次的 API 请求，仅能得到 100 条微博；我们通过标签接口，每一次仅能获取一个特定用户的标签列表。因此，获取大量标签相关数据进行分析，具有一定的难度，也需要时间的积累。

因为权限问题，我们采取这样的策略来获取数据：通过已经创建的微博应用，获取使用微博接口的权限；访问一次 API 后稍作休息，进行下一次访问。

## 2.3 标签相关数据分析

### 2.3.1 数据集

标签相关数据的获取是在 2012 年 2 月到 3 月期间进行的。“statuses/public_timeline”是新浪微博 API 开放的实时公共微博接口，通过该接口，我们可以抽取到此刻刚刚发布出来的无隐私设置的微博信息及其作者信息，因此也具有随机性。从该接口我们一共获取了 1438076 个微博用户的个人信息，随后，通过用户标签接口获取了他们的标签信息。

### 2.3.2 标签的若干特征

（1） 微博用户使用标签功能的情况

据 DCCI（DATA CENTER OF CHINA INTERNET）2010 上半年中国互联网调查数据显示，微博用户经常使用标签功能的人数比例约为 20.2%，而使用功能前三位是评论、关注、转发，分别占 69.8%、60.9%、57.2%。根据采集的数据，我们进一步分析了用户对标签功能的使用情况，如表格 2-3 所示。

统计将在我们获取的 140 多万人的采样集合中进行。从表 2-3 中，可以看

到，采样集合中，44.81%的人（人数超过 64 万）未给自己贴任何的标签，64.73%的人贴的标签数少于 5 个，为自己贴满新浪规定的至多 10 个标签的人仅占 10.2%。用户为自己贴标签的比例依旧较低。

另外，我们从这 140 万用户的采样集合中，分析了在近两年内注册的微博用户的标签使用情况，如图 2-3 所示。我们可以看到，较少使用标签功能的人数依然占据比较大的比例。而注册时间越久，越有可能为自己打上标签。

因此，鉴于标签功能稍低的使用比例，标签功能具有可发展的前景。由于用户自己书写标签还是需要耗费一定的时间，因此如何准确地识别用户兴趣，为用户自动生成、推荐标签是有意义的。而我们首先可以，针对最近注册的用户，提供这样的标签推荐服务。

表 2-3 标签功能使用情况

| 标签数量 | 人数 | 比例(%) | 分段比例(%) |
|---|---|---|---|
| 0 | 644352 | 44.81 | |
| 1 | 84333 | 5.86 | |
| 2 | 57734 | 4.01 | **64.73** |
| 3 | 67741 | 4.71 | |
| 4 | 76772 | 5.34 | |
| 5 | 80792 | 5.62 | |
| 6 | 75871 | 5.28 | |
| 7 | 69478 | 4.83 | |
| 8 | 66292 | 4.61 | 35.27 |
| 9 | 66910 | 4.65 | |
| 10 | 147801 | 10.28 | |

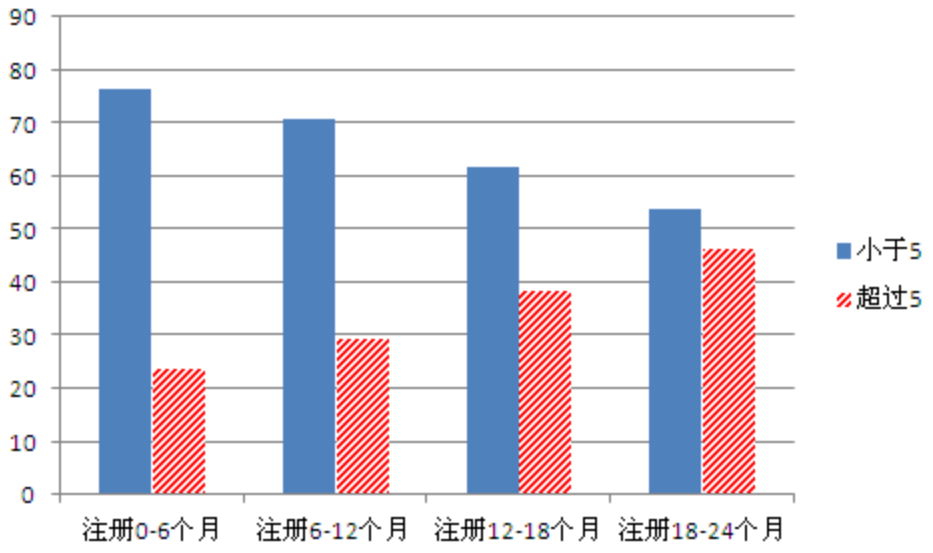



图 2-3 用户注册微博时间与标签使用的关系

（2） 微博用户使用标签功能和活跃度的关系

一般来说，我们可以把一个用户的微博、关注、粉丝数目当做该用户是否活跃的判断因素。我们猜想，用户越活跃，他使用标签功能的可能性是否也就越大呢？

统计依旧在 140 多万人的集合中进行。图 2-4 阐述了这一关系。我们对采样的新浪微博用户的微博、关注、粉丝、标签数目进行了分析。图中微博、关注、粉丝数目取平均值。我们发现数据结果与我们的猜想是一致的，倘若一个用户的活跃度越高，那么他使用标签的情况也就越活跃。

微博活跃用户的微博、关注、粉丝数目往往较多，这样使得微博用户标签自动生成可利用的资源也就更加丰富，识别的用户兴趣越准，生成的标签质量也就可能越高。同时，针对他们的广告推送也就越精准，而活跃用户可能更有机会看到对其推送的广告等，因此，也更具商业价值。

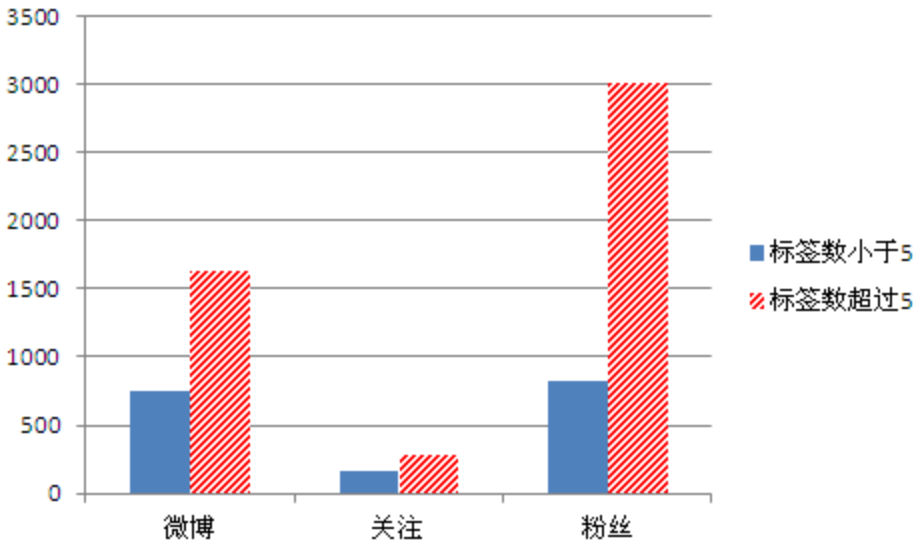



图 2-4 微博用户使用标签功能与活跃度的关系

（3） 标签的分布及语言特征

我们抽取微博用户的标签列表，经过繁简转换、去重后得到 986664 种不同的标签。

**标签的分布规律：**

我们将获取的标签按照使用频数从高到低排序，分布形式如图 2-5 所示，x 轴代表标签序号，y 轴代表使用频数，其中 x 轴轴线还将继续延伸至序号 986664，其 y 值将长期维持在 1。可以看出，微博用户标签基本符合长尾分布，使用频数超过 10 的标签仅占标签总数的 3.94%，而使用频数等于 1 的标签（即“尾标签”）占据了 79.40%。这说明除去极少数的大众化的热门标签外，微博用户给自己添加的标签相当的个性化。这或许也意味着微博用户标签质量可能参差不齐，需要规范。

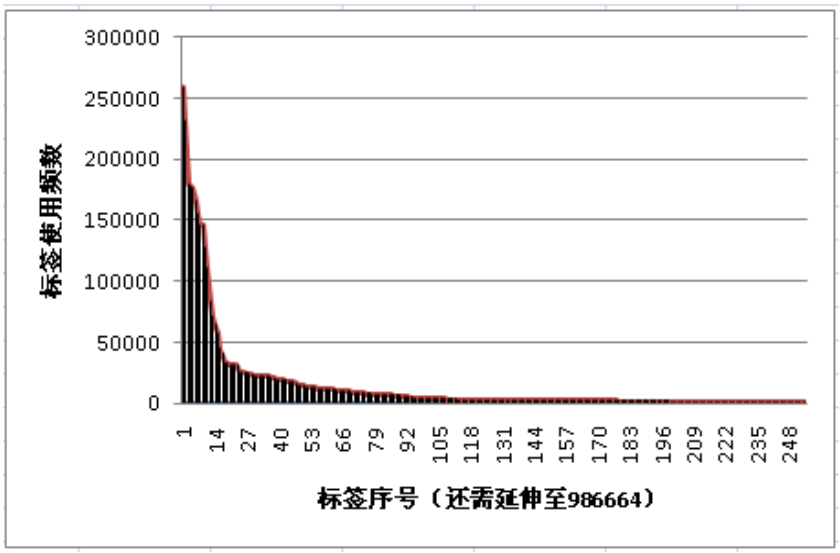



图 2-5 标签的分布

**标签的长度：**

我们对微博用户标签进行中文分词（未去停用词），统计得到标签以词为单位的长度分布如图 2-6 所示。我们可以看到，自定义的用户标签通常是以比较简短的词组形式出现，长度多数在 5 以内。同时，我们注意到，以单个词语作为标签的比例也超过了 10%。

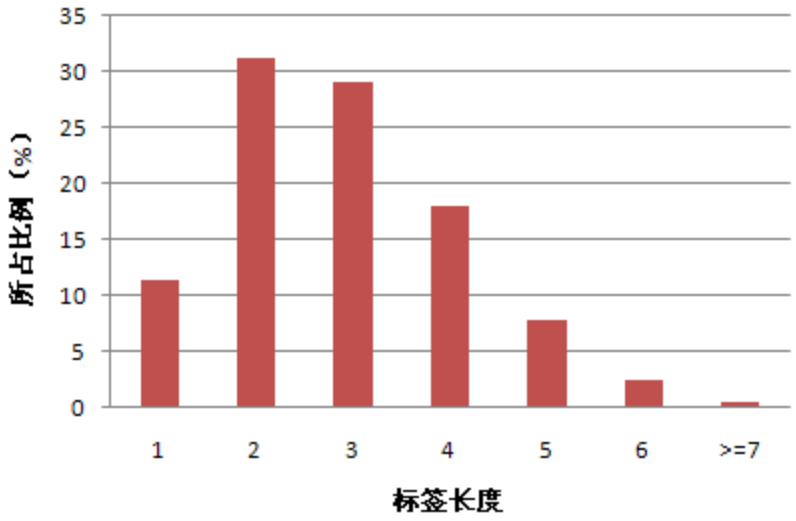



图 2-6 标签长度分布

**标签的中英文分布：**

我们的研究工作是针对中文微博展开的，在获取的 986664 种不同的标签中，我们发现含有英文的标签占 15.42%，纯中文的标签占 84.58%。

**标签的语言特征：**

通过对获取的标签的观察，我们发现，用户多是从这几个维度为自己打标

签：兴趣爱好、职业领域以及对自我性格、形象等方面的描述例如善良、阳光、懒等等。自我描述型标签往往是我们很难挖掘到的，因此，本课题对标签自动生成这一任务的定义也仅限于前两者。

值得注意的是标签中网络流行用语的使用也是比较频繁的，例如大叔控、控苹果等表达对一类事物的痴迷、喜爱的词语。

图 2-7 显示的是使用频率在前 100 的标签及其相应的使用频数。我们大致可以总结出以下一些规律：

这些标签多数仅由长度为一的词语构成；

它们多数都能够体现用户的兴趣爱好或者职业领域特征，例如动漫、篮球、股票、电子商务等；

在这 100 个标签中，我们也可以找到各大门户网站导航中具有明显类别信息的词语，例如：旅游、体育、汽车、互联网、校园、数码、财经等。同时，我们也可以观察到，这些标签缺乏比较明显的类别层次，语义重叠较为严重。如体育、运动、篮球、nba，旅游、旅行等等。

| | | | |
|---|---|---|---|
| 1 音乐 258940 | 26 交友 26061 | 51 偶稀饭睡觉觉 14004 | 76 美女 8396 |
| 2 电影 231895 | 27 搞笑 25106 | 52 金牛座 13396 | 77 历史 8318 |
| 3 美食 180028 | 28 天秤座 24804 | 53 善良 13385 | 78 囧 8087 |
| 4 90后 178437 | 29 水瓶座 23968 | 54 nba 13368 | 79 浪漫 8052 |
| 5 80后 176299 | 30 运动 23194 | 55 创业 13059 | 80 偶喜欢玩 7774 |
| 6 旅游 165836 | 31 网购 22810 | 56 一些事一些情 12527 | 81 媒体 7648 |
| 7 听歌 155840 | 32 艺术 22778 | 57 摄影爱好者 12452 | 82 传媒 7550 |
| 8 旅行 146267 | 33 乐观 22558 | 58 平常心 12440 | 83 网络 7525 |
| 9 时尚 146182 | 34 射手座 22529 | 59 阅读 12160 | 84 文字 7245 |
| 10 宅 128422 | 35 吃 22339 | 60 微博 11958 | 85 营销 7163 |
| 11 自由 111025 | 36 双鱼座 22267 | 61 电子商务 11705 | 86 数码 6964 |
| 12 学生 84407 | 37 双子座 21113 | 62 互联网 11684 | 87 体育 6825 |
| 13 睡觉 71775 | 38 处女座 20666 | 63 宅男 11279 | 88 广告 6678 |
| 14 摄影 65333 | 39 巨蟹座 20348 | 64 羽毛球 10837 | 89 紫色 6646 |
| 15 上网 57582 | 40 足球 19461 | 65 新闻 10664 | 90 淘宝 6016 |
| 16 动漫 44933 | 41 美剧 19449 | 66 爱情 10662 | 91 玩儿 5731 |
| 17 唱歌 41433 | 42 吃货 19424 | 67 文学 10569 | 92 购物 5575 |
| 18 篮球 34092 | 43 大学生 19065 | 68 八卦 10525 | 93 写作 5558 |
| 19 幽默 33026 | 44 生活 18733 | 69 英语 10155 | 94 海贼王 5147 |
| 20 天蝎座 32192 | 45 设计 18354 | 70 我爱uc浏览器 9181 | 95 语录 4924 |
| 21 宅女 32095 | 46 游戏 18292 | 71 财经 8996 | 96 微博控 4861 |
| 22 娱乐 31431 | 47 摩羯座 16096 | 72 it 8888 | 97 校园 4630 |
| 23 看书 31328 | 48 白羊座 15686 | 73 好性格 8600 | 98 大学 4611 |
| 24 小说 26473 | 49 汽车 15138 | 74 手机 8590 | 99 好书 4554 |
| 25 狮子座 26433 | 50 读书 14601 | 75 股票 8485 | 100 星座 4517 |

图 2-7 使用频率前 100 的热门标签及其使用次数

## 2.4 本章小结

本章首先介绍了获取微博用户标签的方法，随后，对获取到的较大规模的用户标签相关数据做出了较为详细的分析。我们发现：1）用户在微博服务中使用标签功能的百分比较低，其使用标签的概率与其活跃度成正比。2）用户的标签集合具有长尾分布的规律，微博用户标签十分个性化，同时也意味着比

较难组织、规范。3）用户标签绝大多数由单个的词语，或者长度较短的短语构成。4）使用频繁的标签多数能够体现用户的兴趣爱好或者职业领域特征，而且往往带有比较明显的类别信息，但是缺乏明显的层次，语义重叠较为严重。

# 第3章 基于文本的标签源分析

## 3.1 引言

本文中，我们不妨把能够生成微博用户标签的资源称为标签源。本文主要着眼于基于内容的用户标签自动生成，因此，也有必要讨论一下在中文微博环境中能够作为标签源的文本。这些文本自然也是通过微博用户相应的行为产生的，它们也通常被称为 UGC[42]（User-Generated Content）。

下面，我们将对产生标签源的行为进行讨论，并且，针对标签源与用户兴趣之间的关系进行进一步的分析。由此，我们也能够确定怎样的标签源生成的用户标签能够较好地体现用户兴趣。

## 3.2 用户产生标签源的行为及统计性质

在中文微博中，能产生文本的行为大致有这两种：用户发布的微博、用户对微博作出的评论。而用户发布的微博中存在原创与转发两种。当用户看到某条感兴趣或者对自己有用的微博时，可能会点击收藏。因此，本文中，我们将可以产生文本作为标签源的行为细分为四种：原创、转发、评论和收藏。

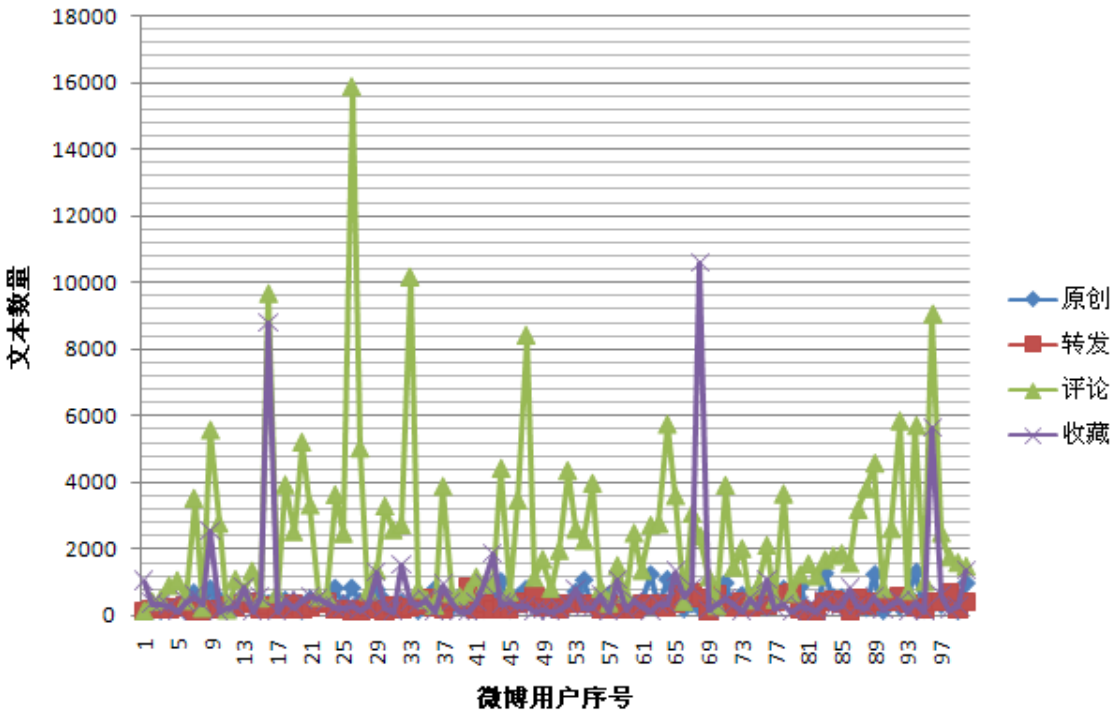



图 3-1 原创、转发、评论、收藏四种行为分布

这四种行为将呈现出怎样的统计性质呢？直觉上，在微博中，有的人偏爱发布自己原创的微博，有的人偏爱转发他人的微博，因此需要对相关数据进行统计，方可得出结论。

本文在我们已经创建过的微博应用“围脖庞统”[2]中，随机抽取了 100 个微博用户，观察这四种行为产生的文本数量，如图 3-1 所示。我们可以看到，原创、转发、收藏这三种行为，不同用户使用的频率不同，在该图中观察不到明显的规律。但是，对于我们随机抽取的 100 个用户，基本都满足“评论行为发生的频次明显高于其他三种行为”这一规律（每一次行为的发生，都将产生一条微博文本，因此图中用文本数量来代表行为发生的频次）。

微博用户的评论行为发生频率一般远远高于其他三种行为，这种现象是可解释的。在第 2 章中，我们提到据 DCCI（DATA CENTER OF CHINA INTERNET）2010 上半年中国互联网调查数据显示，微博用户经常使用的功能，位居首位的就是评论。微博作为一种新型的社交性媒体，对朋友微博的评论正是其社交性的一种体现。

## 3.3 标签源的语义相似度

用户每天在微博上通过原创、转发、评论和收藏这四种行为，产生的内容均能体现出自己感兴趣、关注的方面。但是，它们讨论的话题是否趋于一致，在语义上的相似度如何呢？据我们所知，很少有人对这一点展开研究。下面，我们将从词语（term）级别和主题（topic）级别来讨论标签源的语义相似度。

同样，我们从已经创建的微博应用“围脖庞统”中，抽取一部分用户，这些用户通过这四种行为得到的微博文本数量至少超过了 100 条。我们一共得到 105 个符合条件的用户，我们采用“集成策略”，分别将用户每种行为得到的微博文本合成一篇大文档来进行相关的计算。

### 3.3.1 词语级别的相似度

对于每种行为得到的微博文档，我们采用 TFIDF 策略抽取前 n 个权重较大的词语集合，比较它们之间的相似度（TFIDF 策略具体在第 4 章中将有详细介绍）。这里，词语集合的相似度用 Jaccard 系数衡量，如公式(3-1)。即两个集合 A、B 的交集元素在 A 与 B 的并集中所占比例，称为 Jaccard 系数。

$$Jaccard(A,B)=\frac{|A\cap B|}{|A\cup B|} \tag{3-1}$$

我们分别取前 10、20、30……200 个词语集合计算两两之间的相似度，结果如图 3-2 所示。其中，ori 代表原创、ret 代表转发、cm 代表评论、fav 代表

[2] http://pangtong.sinaapp.com/

收藏，这些字符以“_”连接，代表两两间的比较。

此外，我们在由词语构成的向量空间上，用余弦相似度如公式(3-2)来衡量各种文本标签源的语义相似度，结果如图 3-3 所示。

$$\cos(A,B)=\frac{\sum_{k=1}^{n}(w_{k,A}\times w_{k,B})}{\sqrt{\sum_{k=1}^{n}w_{k,A}^{2}}\times\sqrt{\sum_{k=1}^{n}w_{k,B}^{2}}} \tag{3-2}$$

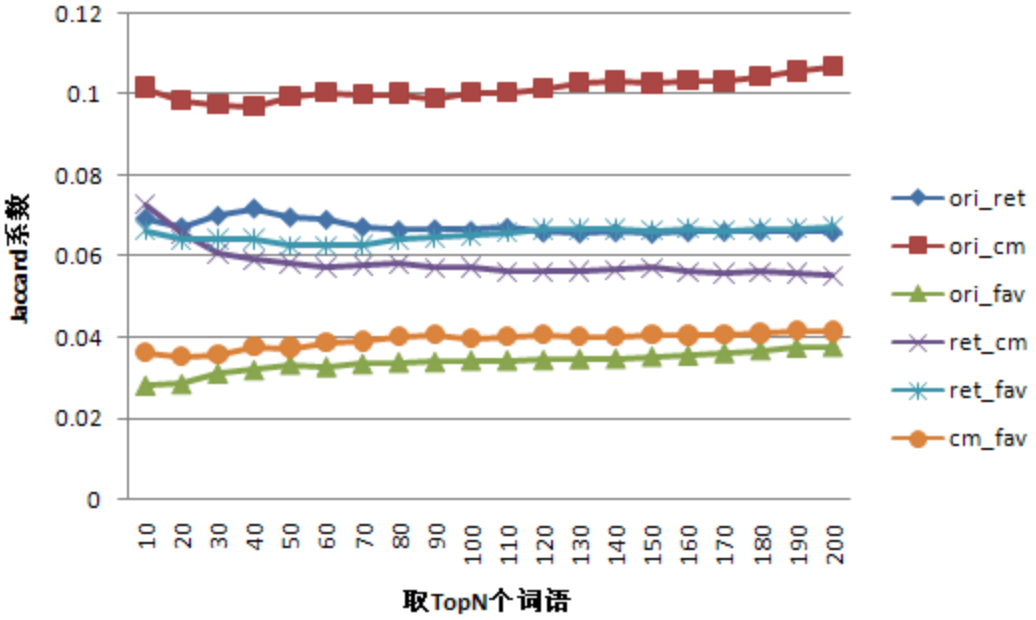



图 3-2 以 Jaccard 系数衡量的词语级别的相似度

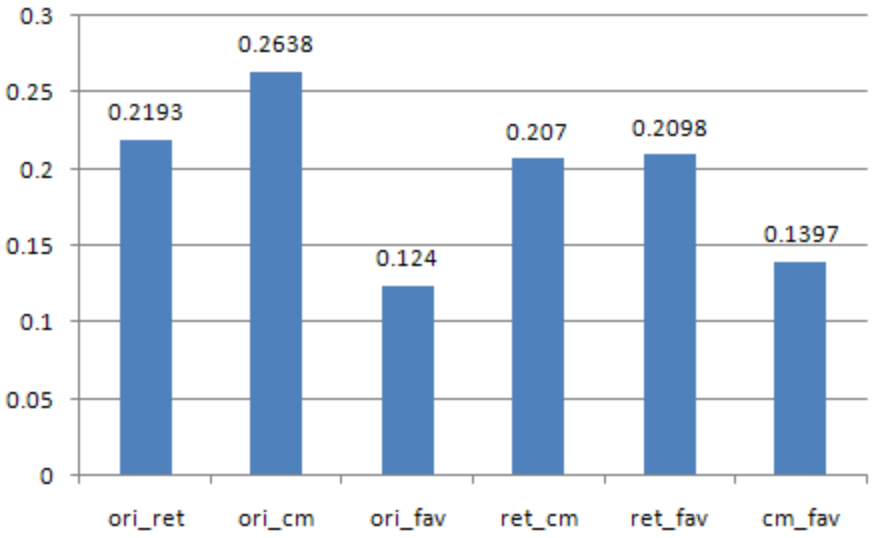



图 3-3 以余弦相似度衡量的词语级别的相似度

我们发现，在词语级别上，各种文本标签源之间的语义相似度较低。其中，原创和收藏的相似度最低，而原创和评论的的相似度相对较高。

### 3.3.2 主题级别的相似度

主题模型是近年来比较流行的对文本隐含主题进行建模的方法，例如 Latent Dirichlet Allocation（LDA）。LDA 可将传统向量空间中词的维度转变为 N 个 topic 的维度（N 是我们自定义的），也是一种有效的克服文本稀疏的降维手段。LDA 的原理我们将在第 5 章展开具体的介绍。

在这里，我们用 LDA 对四种微博文档进行建模。在由 n 个 topic 组成的主题向量空间上（这里设置为 200），计算文档间的相似度，以余弦相似度衡量，如公式(3-2)。计算结果如图 3-4 所示。

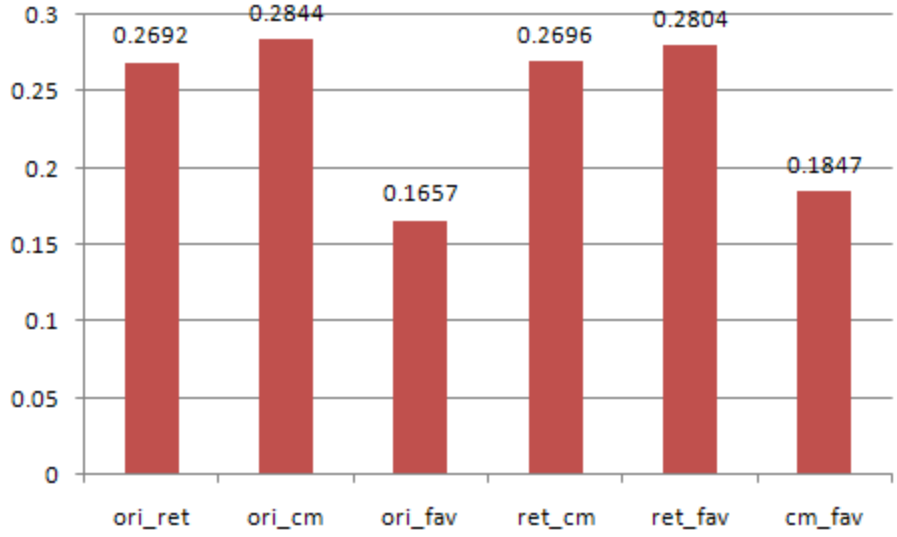



图 3-4 主题级别的相似度

综合图 3-2、图 3-3 和图 3-4 的词语和主题级别的相似度结果，我们可以知道，尽管原创、转发、评论和收藏能在不同程度上反映用户的兴趣，但是它们产生的微博文本讨论的话题还是不大一致的，反映在结果上就是词语和主题级别的相似度都比较低。

由图 3-2 和图 3-3 我们得到四种微博文本在词语级别上的相似度由高到低排列为：原创与评论、原创与转发、转发与收藏、转发与评论、评论与收藏、原创与收藏；由图 3-4 我们得到四种微博文本在主题级别上的相似度由高到低排列为：原创与评论、转发与收藏、原创与转发、转发与评论、评论与收藏、原创与收藏。虽然两者的排列在“原创与转发、转发与收藏”处有稍许差别，但是从这些图中，我们可以看到它们的相似度是相当逼近的，因此，相似度规律在词语和主题级别基本一致。

“原创与评论”体现出相对较高的相似度，而“原创与收藏”相似度最

低。我们可以想象，用户通过原创与评论产生的微博文本，由于都是用户自己编辑发布的，是最接近用户语言风格的，因此，结果也是可以解释的。

## 3.4 标签源选择对反映用户兴趣的影响

原创、转发、评论和收藏能在不同程度上反映用户的兴趣，但是，哪种行为的贡献最大，需要更加深入的考察。

### 3.4.1 方案

我们将四种行为得到的微博文档进行用户兴趣建模，通过考察它们对微博推荐的质量来衡量反映用户兴趣的程度大小。模型将用基于词语的向量空间模型和经过 LDA 降维的主题向量空间模型构建。

具体来说，对于某用户接收到的微博文本，我们将逐一考察它们与用户兴趣模型的相似性，按照相似程度由高到低重新排序，推荐给该用户。排名越靠前的，我们认为越符合用户的口味。相似性依旧用余弦相似度来衡量。

我们请了 6 位标注者，获取了他们在 2011 年 12 月最后两周能够看到的所有微博。他们将对自己感兴趣的微博做出标记。整个标注过程是可视化的，在我们创建的应用微博选读[3]上进行，当用户看到感兴趣的微博时，将点击红心按钮，而系统将会把这一标注行为记录下来，标注界面如图 3-5 所示。由于我们每天接收到微博信息量很大，标注者通过我们的系统可能只标注到部分感兴趣的内容，但是，这足以进行我们的比较实验。


图 3-5 标注界面

[3] http://xuandu.hit.edu.cn

### 3.4.2 数据

表格 3-1 显示了 6 位标注者的原创 ori、转发 ret、评论 cm 和收藏 fav 的微博数量。它们将用于构建兴趣模型。

表 3-1 标注者微博数量

| 用户 ID | #ori | #ret | #cm | #fav |
|---|---|---|---|---|
| 1830105817 | 32 | 143 | 191 | 130 |
| 1772689277 | 145 | 331 | 656 | 587 |
| 1648851542 | 62 | 108 | 212 | 63 |
| 1143993665 | 95 | 278 | 145 | 46 |
| 1645017212 | 161 | 387 | 502 | 41 |
| 1083622107 | 86 | 100 | 155 | 69 |

表格 3-2 是他们在 2011 年 12 月最后两周接收到的微博数量以及他们标注为“感兴趣”的数量。

表 3-2 标注微博数量

| 用户 ID | #接收微博 | #感兴趣 |
|---|---|---|
| 1830105817 | 1602 | 111 |
| 1772689277 | 1955 | 129 |
| 1648851542 | 8493 | 177 |
| 1143993665 | 2261 | 105 |
| 1645017212 | 1346 | 115 |
| 1083622107 | 2430 | 50 |

### 3.4.3 结果

我们采用公式(3-3)的 P@N 指标来衡量推荐的效果，考察了标签源的不同组合反映用户兴趣的程度。图 3-6、图 3-7、图 3-8 分别体现的是单种标签源、任意两种标签源组合、任意三种标签源组合构建兴趣模型，得到的推荐效果。

$$P@N = \frac{\text{前N条微博中用户感兴趣的微博数目}}{\text{N}} \tag{3-3}$$

我们可以看到，基于词语和基于隐含主题计算相似程度排序得到的推荐效果规律基本一致。

就**单种标签源**来说，转发和收藏最能体现用户兴趣，原创次之，评论最差。我们猜测原因，对于大多数微博用户来说，原创和评论更倾向于日常生活琐事的叙说。

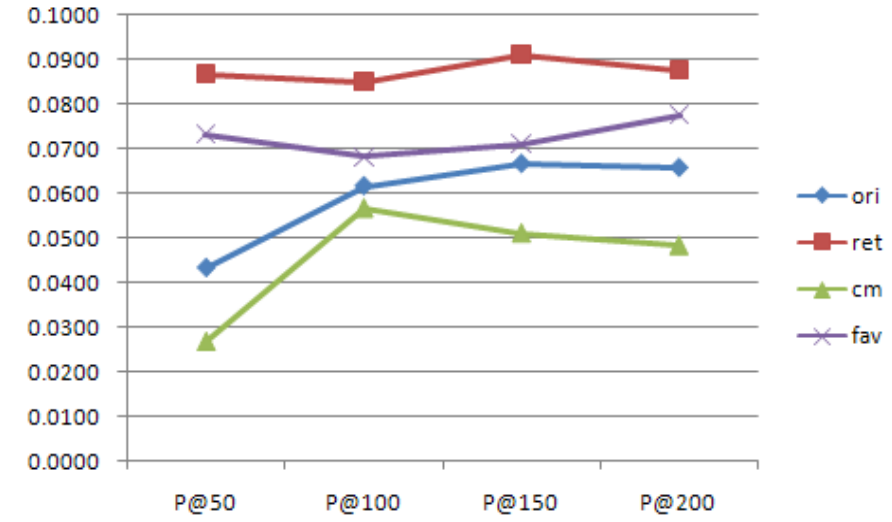
就**标签源两两组合**来说，原创与转发、转发与收藏最能体现用户兴趣，而原创与评论的组合最差。

**任意三种标签源组合**，lackori、lackret、lackcm、lackfav 分别代表除去原

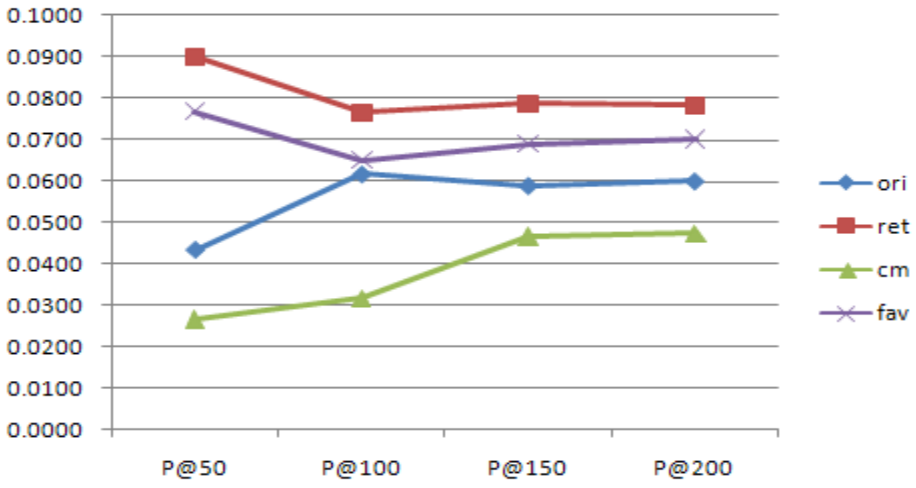
创、转发、评论和收藏的标签源组合。lackcm 的组合效果最好，再一次证明了，评论反映用户兴趣的程度稍差。

我们取每种组合中表现最好的类型，与四种标签源共同构建的兴趣模型（用 all 表示）进行比较，如图 3-9。转发的表现依旧最好，其次是原创与转发的组合，再次是除去评论的组合，而四种标签源虽然都能体现用户兴趣，但它们共同构建的模型效果却最差。

分析了四种行为反映用户兴趣的程度后，我们在之后的研究工作中对标签源的选择也就更有了依据。

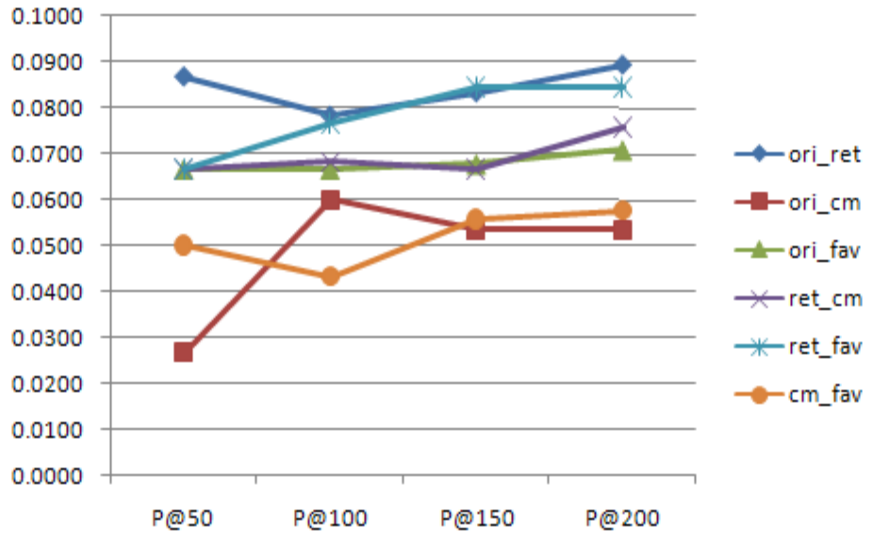



(a) 基于词语

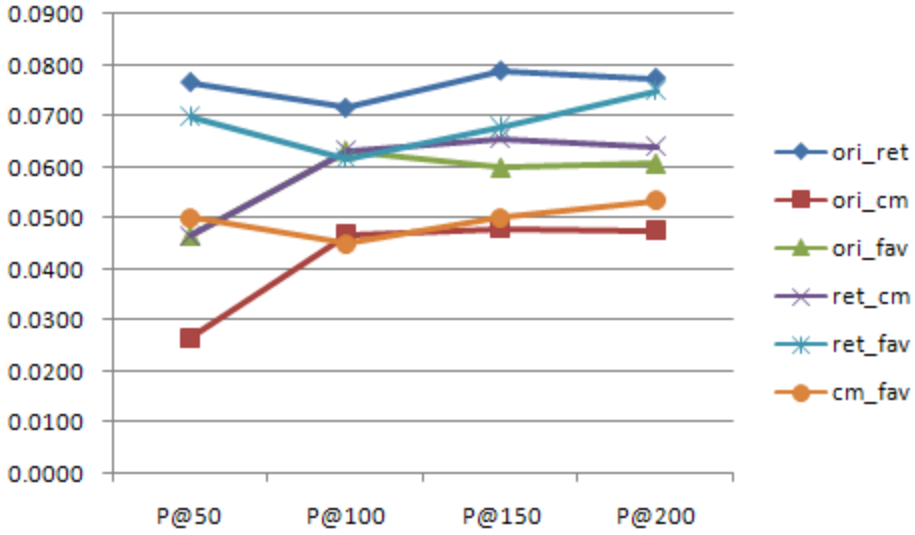



(b) 基于隐含主题

图 3-6 单个标签源构建兴趣模型

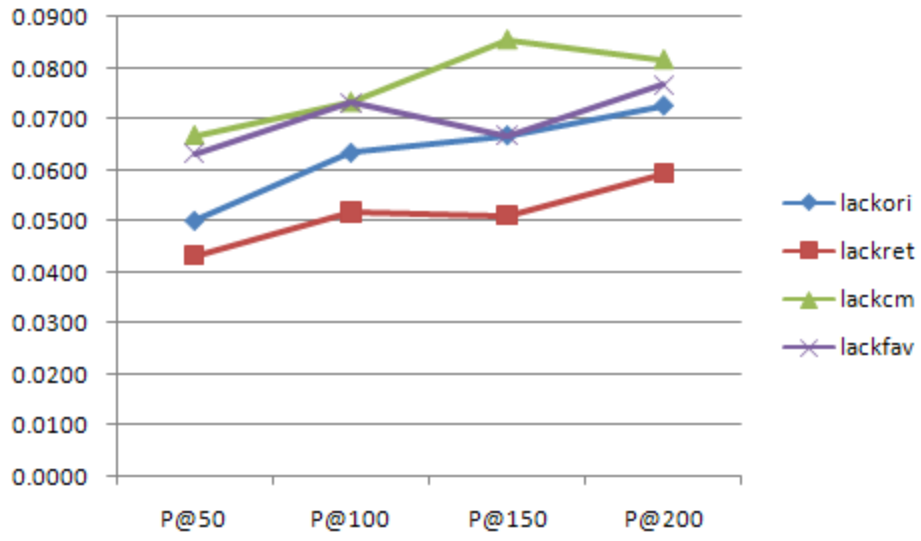



(a) 基于词语

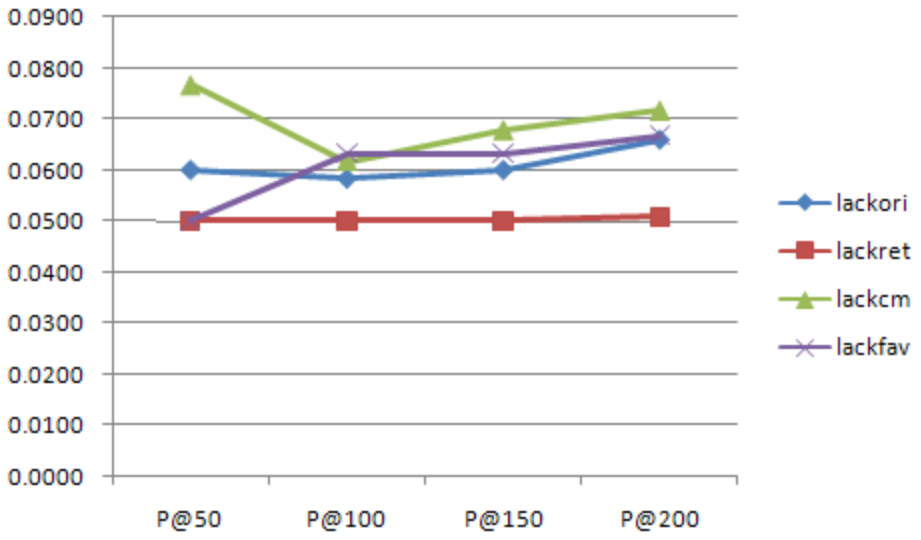



(b) 基于隐含主题

图 3-7 标签源两两组合构建兴趣模型

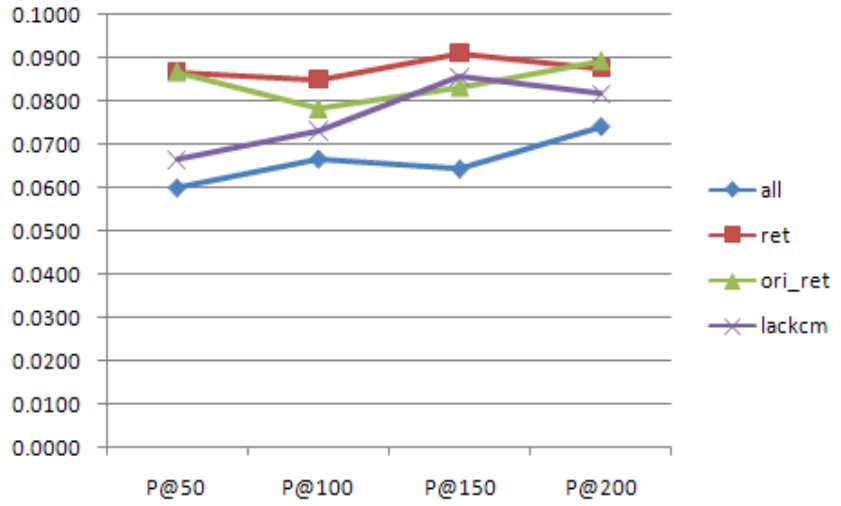



(a) 基于词语

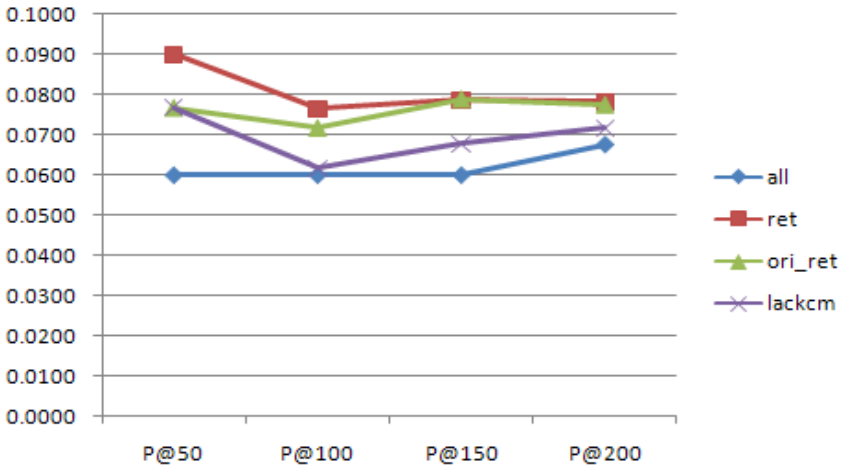



(b) 基于隐含主题

图 3-8 任意三个标签源组合构建兴趣模型




(a) 基于词语




(b) 基于隐含主题

图 3-9 每种标签源组合中表现最优的组合进行比较

## 3.5 本章小结

本章对能够生成微博用户标签的文本即基于文本的标签源进行了较为深入的分析，分析主要从能够产生文本标签源的行为、标签源的语义相似度以及标签源体现用户兴趣的程度这三个方面进行。我们发现，1）标签源之间的语义相似度较低，相对而言，原创与评论的相似度最高，而原创与收藏之间的相似度最低；2）四种文本标签源及其组合中，转发最能体现用户兴趣，原创与转发的组合以及收藏都有不错的表现，而评论最差。

# 第4章 基于关键词的标签自动生成

## 4.1 引言

关键词一般是由单个词语或者多个词语组成的短语。通过第 2 章中，对真实微博用户标签的观察，我们发现多数标签是由单个词语或者长度较短的短语构成。直觉上，关键词也是能够反映出用户的兴趣的。因此，本章中，我们将尝试不同的方法，从关键词自动生成的角度为微博用户添加标签。

在本章中，自动生成用户标签的流程大致是这样的：从用户微博文本中获取候选关键单词，通过某种算法排序得到单个的关键词，再基于规则将它们进行一定的扩展作为用户的标签。

从第 3 章，我们知道，用户原创和转发产生的微博文本对反映用户兴趣的贡献仅稍微次于仅由转发产生的文本。微博 API 提供的用户发布的微博列表 UserTimeline 既包含原创又包含转发文本。而且，我们相信，用户在原创微博中经常提及的事物也能作为反映用户兴趣的标签。出于上述两点考虑，我们的标签源选择了原创与转发组合的微博文本，即用户自己发表的微博。

## 4.2 基于 TextRank 的生成方法

### 4.2.1 TextRank 概述

Rada Mihalcea 等人[43]于 2004 年提出了 TextRank 算法用于文本关键词抽取。这是一种类似于 PageRank[44]的图模型算法。

TextRank 将文档中的词语类比于互联网网页，而词与词之间的联系类比于网页之间的链接关系。也就是说，算法认为文本是一个由词语构成的网络或者说是一个由词语作为节点构成的图，词之间的语义关系构成边。在图中越重要的词，也就越可能是关键词。

形式化地，我们令$G=(V,E)$代表文本中由词语构成的有向图，$V$ 为词语节点，$E$ 为边。对于每一个节点$V_i$，$In(V_i)$ 代表指向它的节点集合，$Out(V_i)$ 代表节点$V_i$指向的节点集合。$w_{ij}$ 代表$V_i$和$V_j$之间边的权重。我们确定一个滑动的文本窗口，窗口中包含 k 个词，倘若两个词语同时出现在这个窗口中，我们可以称它们共现。可以将词对间的共现次数作为连接它们的边的权重。节点$V_i$的分数计算如公式(4-1)所示。

$$S(V_i)=(1-d)+d*\sum_{V_j\in In(V_i)}\frac{w_{ji}}{\sum_{V_k\in Out(V_j)}w_{jk}}S(V_j) \tag{4-1}$$

从公式(4-1)，我们可以看出，这是一个迭代的过程。d 为 0 到 1 之间设定

的一个值，代表从给定的一个节点跳向图中随机的一个节点的概率。当然，在实际的文本处理中，图也可以是无向的；节点也可以用字等来表示。

其实，TextRank 的主要思想就是，一个词的重要性由其他与其关联的词决定。

TextRank 的算法流程大致是这样的：

1）确定文本的最佳代表形式：词语或者单个字或者其他，并将其作为图中的节点；

2）构建节点之间的边，例如共现信息当做权重；

3）迭代，直到算法收敛；

4）将节点按照分数排序，得到关键词。

### 4.2.2 生成方法

在该方案中，我们将借助于 TextRank 算法，为微博用户自动生成标签。方案的流程如图 4-1 所示。

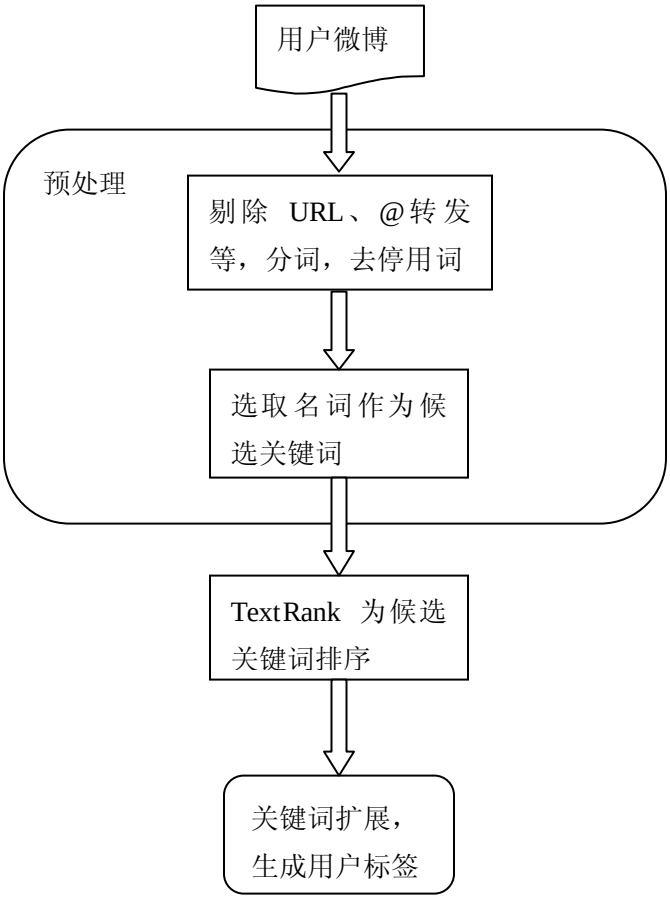



图 4-1 基于 TextRank 自动生成标签流程

1） 预处理

当得到用户发布的微博文本之后，我们采取“合成”策略，即将用户发布的所有微博合并成一个大文档进行处理。为避免不必要引入不必要的噪声，我们先去除微博文本中带有的 URL 链接，以及文本中@某某人的字样。分词后，去除预定义的停用词。

在很多的研究中，发现文档的关键词都倾向于名词性质的词，因此，在本文中，我们的候选关键词也选择了名词。

2） TextRank 排序

预处理后，我们为用户微博文本构建以候选关键词（即选取的名词）为节点的无向图。我们的滑动窗口定为一条微博的长度，即倘若两个词在同一条微博中出现，我们就认为它们之间存在较强的语义联系，共现次数加 1。我们对每一条微博进行同样的词对共现次数提取。随后，图节点间边的权重记为它们在该用户微博文本中的共现次数。

图构建完之后，我们开始计算每个名词节点的分数，对公式(4-1)进行适当调整，无向图的计算公式如(4-2)所示。其中，$E(V_i)$ 表示与节点 $V_i$ 连接的所有节点集合。

$$S(V_i)=(1-d)+d*\sum_{V_j\in E(V_i)}\frac{w_{ji}}{\sum_{V_k\in E(V_j)}w_{jk}}S(V_j) \tag{4-2}$$

常数 d 通常被设置为 0.85。分数计算是个迭代的过程，直至收敛结束。由于我们关注的是关键词的排序，因此，当所有词排列顺序不变时，我们将停止计算。

3） 关键词扩展

由 TextRank 排序得到的都是单个的名词，可能不足以表达用户的兴趣。因此，我们将基于规则，以名词为中心，扩展成关键词串，作为标签。

文献[45]发现，大部分的关键词都是名词性词组。目前，进行扩展得到关键词（多个词语组成）的方法大致有两种[43]，一种是进行词性标注后，抽取符合一定模式的词串，例如形容词与名词的组合；另一种是一种“后处理”方法，得到单个的关键词排序列表后，查看排名前 N 的单词，倘若它们在文档中相邻，则进行组合。

考虑到中文词性标注的准确率还不是很高，我们采取第二种方法进行扩展：查看 TopN 的单个关键词 TextRank 排序列表，是否有相邻的组合存在。考虑到组合是否稳定的问题，我们仅抽取出在原文中出现次数超过 3 次的组合，紧接着计算权重。扩展后的词串权重，为组成它的词语的 TextRank 分数之和。按照权重排序后，抽取前 10 作为自动生成的用户标签。从扩展的过

程，我们可以看出，生成的用户标签或者是词语或者是由词语组成的词串。

### 4.2.3 实验与结果分析

#### 4.2.3.1 实验数据

1） 测试用户

本文主要研究的是基于微博内容的用户标签自动生成，因此我们从已创建的微博应用“围脖庞统”中随机选择了 40 位已发布过微博文本的用户作为我们的测试用户。他们发布的平均微博条数为 407.05 条，具体分布如图 4-2。后续的所有方案也将使用这些测试用户。

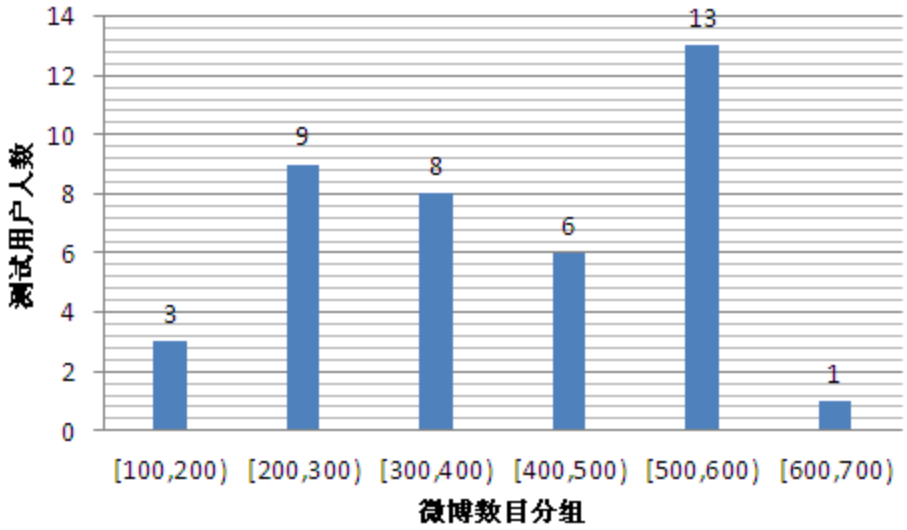



图 4-2 测试用户微博数目分布

2）Baseline

我们采用 TFIDF 为候选关键词排序的策略作为我们的 Baseline 系统。排序方式如公式(4-3)所示。

$$\text{tfidf(t,u)}=\text{tf(t,u)}\times\log(\frac{U}{U_t}) \tag{4-3}$$

其中，tf(t,u)表示用户 u 的微博文本中词 t 的频率，$U$ 表示微博语料中用户的总数，$U_t$ 表示微博文本中包含词语 t 的用户数。

$$\text{tfidf}(t)=tf(t)\times\log\frac{N}{df(t)} \tag{4-4}$$

公式(4-3)实际就是我们常见的 TFIDF 公式(4-4)的变形。式(4-4)中，$tf(t)$ 代表某篇文档中词语 t 的频率，N 代表语料的文档总数，$df(t)$ 表示语料中包含词 t 的文档数目。既然我们的微博文档是由用户发布的所有微博合成的，那么排序公式(4-3)的描述是更为贴切的。

抽取完关键词后，我们按照同样的方式进行扩展，生成用户标签。

我们从第 2 章获取的用户群中随机爬取了 23430 人的微博，他们具体的微博数目分布如图 4-3 所示。相应的 TFIDF 信息由此计算产生。

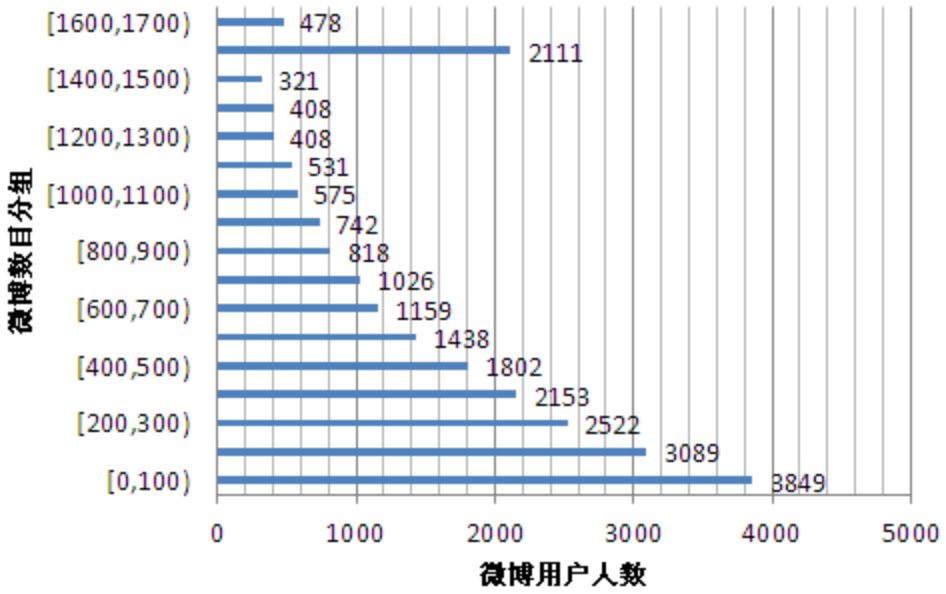



图 4-3 TFIDF 策略微博语料分布

#### 4.2.3.2 评价方法

在第 2 章中，我们发现用户为自己添加的标签中包含部分自我描述型的内容，例如“闷骚”、“善良”、“没完没了的任性”等，并且那些普遍型的标签例如“音乐”大部分用户都有添加，可能并未深入地体现用户的兴趣。本文专注于利用用户自己产生的微博内容进行标签的自动生成，生成的标签未必包含在当前用户的标签集合中。因此，我们采用人工评价的方式。

评价者浏览测试用户的所有微博内容，借助于用户已经为自己添加的标签等，进行评定。评定准则有两条：一是我们生成的结果是否能体现用户的兴趣；二是我们生成的结果是否适合作为用户标签。我们为用户自动生成 10 个标签，按照方法中的权重计算方式进行排序。

评价指标采用信息检索领域经典的评价指标 P@N，表示生成的前 N 个用户标签结果的准确率，公式如(4-4)。

$$P@N = \frac{\text{前N个标签结果中生成正确的数目}}{\text{N}} \tag{4-4}$$

由于评价过程比较主观，因此将有两位评价者分别对生成的标签结果作出评定，并使用 kappa 值比较他们评定的一致性。

#### 4.2.3.3 结果与分析

我们按照基于 TextRank 的方法对 40 位测试用户进行了标签自动生成，同时与我们的 Baseline 方法进行对照。结果如表 4-1 所示。

其中，TextRank-1 在构建词语同现网络时，是将滑动窗口设置为整条微博；而 TextRank-2 在构建网络时，当遇到转发微博，我们将滑动窗口设置为用户原创内容或者用户转发内容。

例如，“*谷歌卖人品，做担保生意了。符合时代发展的需要* //@新浪科技：谷歌今天推出“诚信商家”（Google Trusted Stores）免费认证项目，曲线渗透电商领域。符合一定资质的商家可以申请加入，并获得一个徽章；消费者在认证商家购物时如果遭遇欺诈，谷歌会为其撑腰，代表消费者与商家交涉，必要时甚至提供全额退款~http://t.cn/zOFKxsR 谷歌简直是活雷锋啊~”。

在上述微博中，我们分别将字体不同的内容视为两个滑动窗口，我们猜测用户原创内容(斜体字)和真正转发的内容，语义间的联系稍弱。

表 4-1 基于 TextRank 标签生成结果(%)

a) 评价者 A 的评价结果

| | **TFIDF** | **TextRank-1** | **TextRank-2** |
|---|---|---|---|
| P@1 | 55.00 | **62.50** | 60.00 |
| P@2 | 48.75 | **56.25** | 55.00 |
| P@3 | 41.67 | 47.50 | **50.83** |
| P@4 | 40.63 | **50.63** | 48.75 |
| P@5 | 37.50 | **47.50** | 44.00 |
| P@6 | 35.42 | **45.42** | 43.33 |
| P@7 | 32.50 | **44.29** | 42.86 |
| P@8 | 29.69 | **42.50** | 40.94 |
| P@9 | 29.44 | **40.28** | 38.89 |
| P@10 | 28.25 | **38.25** | 37.75 |

b) 评价者 B 的评价结果

| | **TFIDF** | **TextRank-1** | **TextRank-2** |
|---|---|---|---|
| P@1 | 50.00 | **62.50** | 55.00 |
| P@2 | 43.75 | **52.50** | **52.50** |
| P@3 | 40.83 | 45.83 | **50.00** |
| P@4 | 38.13 | **48.13** | 47.50 |
| P@5 | 36.50 | **48.50** | 46.00 |
| P@6 | 36.25 | **47.50** | **47.50** |
| P@7 | 33.93 | **44.64** | 44.29 |
| P@8 | 31.88 | 42.50 | **42.81** |
| P@9 | 31.39 | **41.11** | 40.00 |
| P@10 | 29.75 | **40.25** | 38.75 |

对于每一种生成标签的方式，我们计算了两位评价者的评价结果一致性，用 Cohen's kappa 系数衡量，见表 4-2。kappa 值在-1 到 1 之间，值越大一致性越好。一般，我们认为 kappa 值超过 0.75，则评定的一致性是极好的；kappa 值在 0.40 与 0.75 之间，一致性是比较不错的；kappa 值低于 0.40，一致性比较差。

表 4-2 评价者 A、B 对标签生成结果评定的一致性

| | TFIDF | TextRank-1 | TextRank-2 |
|---|---|---|---|
| Cohen's kappa | 0.5770 | 0.6227 | 0.6084 |

从表 4-2 可以看出，两位评价者对标签生成结果的评定一致性是比较不错的，而表 4-1 中，两份评价结果所体现出的规律是基本一致的，因此，我们相信基于此得出的结论将比较可靠。

从结果来看，我们基于 TextRank 的方法生成用户标签的表现要优于 baseline，在 Top10 上平均高出了近 10 个百分点。而使用整条微博构建出来的词语网络 TextRank1 效果是优于 TextRank2。这说明，转发微博中，用户原创的部分与转发的部分，语义联系还是比较强。这可能主要是因为，对于大部分转发微博来说，两个部分的内容所讨论的话题是基本一致的。

## 4.3 基于聚类分析的生成方法

从 4.2 节“基于 TextRank 的生成方法”中，我们发现自动生成的标签能够从一定程度上反映用户的兴趣。但是，我们也察觉到，当微博用户某方面的兴趣表现特别明显时，用 TextRank 这种基于统计的方法，将会出现意义相近的标签堆积的现象。因此，我们将尝试基于聚类分析的方法，自动生成用户的标签。

基于聚类分析的方法自动生成标签，简单来说，流程大致如下：

1） 将用户发布的微博文本进行预处理，得到名词作为候选关键词；

2） 计算任意名词对间的相似度，进行聚类分析；

3） 排序聚类簇；

4） 为 Top10 的聚类簇选取代表词；

5） 扩展代表词，得到用户标签。

下面，我们将对涉及的关键技术与原理，自动生成的步骤做具体的阐述。

### 4.3.1 关键技术与原理

#### 4.3.1.1 词语相似度

文献[46]中，对词语相似度（word similarity）与词语相关度（word relatedness）做出了较为详细的讨论。一般来说，词语相关度较词语相似度具有更宽广的范畴。例如，我们可以说“汽车”与“汽油”这两个词语高度相关，但却不具有较高的相似度；而“汽车”与“车轮”两词则具有较高的相似度。我们的方法中，计算的词语相似度严格来说是词语的相关度。但是，我们通常不对着两个术语做严格的区分。

一种计算词汇相似度的方法是基于文档内部的同现关系的。我们确定一个滑动的窗口，倘若两个词语在其中同时出现，则我们认为它们之间具有语义关

联。通过计算共现次数，得到它们的语义相似度。

另一种方法是借助外部电子资源来衡量词语之间的相似度。

文献[47]中，作者提出了 ESA（Explicit Semantic Analysis）的方法来计算任意词语配对或者任意长度文本配对之间的语义相似度。ESA 方法借助了英文维基百科上百万条的词条信息。对每个词语，用该词分别在这上百万条词条文本信息中的类似 tfidf 权重，构成了一个上百万维的向量空间来表示该词。有了这样的向量空间，便可以用传统的相似性度量指标例如余弦相似度来衡量词语之间的相似度。ESA 其实类似于经典的计算词语相似度的分布式方法。因为每一维代表词的维度都是维基百科词条，可读性可解释性相比潜在语义索引 LSA 等方法更好，因此，作者也将该方法命名为 Explicit Semantic Analysis。

由于我们计算词语相似度是用于构建词语间的相似矩阵，为后续的词语聚类做准备。微博文本相当短小，使用文档内部共现的方式，相似矩阵将相当稀疏，将直接导致聚类效果较差。另外，有关文献[15]也通过实验证明，基于文档内部的同现关系计算出的词语相似度并不好。因此，我们选用了第二种方式：借助于外部资源，百度百科。我们爬取了大约 160 万个百科词条的信息，用作语料，利用词语之间的互信息衡量它们之间的距离（相似度），如公式(4-5)。

$$pmi(t_1,t_2)=\log\frac{N*tf(t_1,t_2)}{tf(t_1)*tf(t_2)} \tag{4-5}$$

其中，N 表示语料中词语的总数目，$tf(t_1)$、$tf(t_2)$分别表示词$t_1$、$t_2$在百度百科语料中出现的频次，而$tf(t_1,t_2)$代表两次在语料中相邻出现的频次。

#### 4.3.1.2 聚类技术

我们在这里简要介绍下聚类技术

聚类是一种将数据对象划分成相似的集合（簇）的过程。常见的聚类技术可以分为划分方法例如 K-Means、层次方法例如层次聚类、基于密度的方法、基于网格的方法等等[48]。很多方法需要预先设定 K 值，即预先设置需要聚成多少个簇。

我们的方法中，对词语的聚类实际上是一个探索的过程。我们的目标虽然是自动生成 10 个最适合用户的标签，但这并不意味着我们就预先设置好要聚成 10 个簇。因此，我们选择了层次方法中的自底向上的凝聚层次聚类。首先将每一个词语都当成一个独立的簇，也就是说我们需要对 N 个词语进行聚类，那么初始的时候就有 N 个簇。然后，根据簇之间的距离，选择最近的两个簇逐一合并，直到 N 个词语都聚成了一个簇。经过这种方法，我们可以得

到一颗具有层次的聚类树，在此之上观察哪一层的聚类效果最符合我们的用途。

层次聚类有一个需要注意的问题就是在逐一合并簇的过程中如何度量两个簇$C_1$、$C_2$之间的距离。根据衡量方式的不同，我们可以将层次聚类细分为以下几种方法：

1）单连接算法，也称为最近邻聚类算法。算法使用两个簇中最近对象的距离作为簇间的距离，当这个距离超过一定阈值的时候聚类终止，度量公式如(4-6)所示。其中，对象$o_1$、$o_2$分别属于簇$C_1$、$C_2$。

$$d(C_1,C_2)=\min_{o_1\in C_1,o_2\in C_2} d(o_1,o_2) \tag{4-6}$$

2）全连接算法，也称为最远邻聚类算法。算法使用两个簇中最远对象的距离作为簇间的距离，当这个距离超过一定阈值的时候聚类终止，度量公式如(4-7)所示。

$$d(C_1,C_2)=\max_{o_1\in C_1,o_2\in C_2} d(o_1,o_2) \tag{4-7}$$

3）均值距离/平均距离算法。前两种方法对坏数据都比较敏感，采用式(4-8)或者(4-9)作为簇间的距离是一种折中的算法。其中，$m_1$、$m_2$是两个簇的均值。$n_1$、$n_2$分别是两个簇中对象的数目

$$d(C_1,C_2)=d(m_1,m_2) \tag{4-8}$$

$$d(C_1,C_2)=\frac{1}{n_1 n_2}\sum_{o_1\in C_1}\sum_{o_2\in C_2} d(o_1,o_2) \tag{4-9}$$

### 4.3.2 生成方法

基于聚类分析的标签自动生成，我们将借助层次聚类算法进行，整个生成标签的流程如图 4-4 所示。

1） 预处理

预处理过程与基于 TextRank 的生成方法相同，预处理后我们将得到一个候选的关键词集合，词语的词性都为名词。

2） 词语层次聚类

我们采用的自底向上的层次聚类对词语进行聚类，具体算法如图 4-5 所示。而待聚类的词语集合是 TextRank 权重前 200 的词语集合。我们认为 Top200 的词语集合已经体现了用户的绝大多数兴趣。

在聚类算法开始之前，其实我们需要为词语构造初始距离矩阵。在本方法“关键技术与原理”这一部分我们已经介绍了用词语间的互信息来衡量距离的方法。我们将稍微放宽“相邻”这一条件，当两个词语出现在以 4 为大小的无序窗口中，我们认定这个两个词语是相邻的。

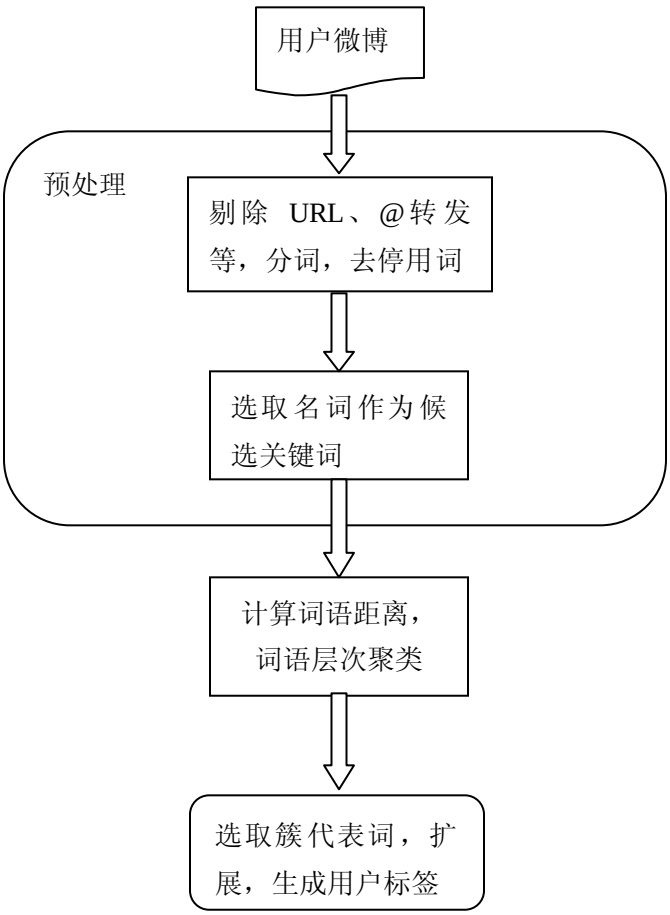



图 4-4 基于聚类分析的标签自动生成流程

3） 选取簇代表词，扩展

通过层次聚类，我们将选定某个聚类效果较好的聚类层。语义相关度较高的词语将被聚成一个簇。通过这样的一个聚类，我们其实可以以探索的态度去发现用户兴趣的大致维度。对于聚类形成的每一个簇，我们需要选择合适的词来代表它。在 4.2 节中，我们已经为每个词语计算了相应的 TextRank 分数。在本方法中，我们选用簇中拥有最高 TextRank 分数的词语作为簇代表词。

选取完簇代表词后，我们按照基于 TextRank 生成方法中同样的策略对词进行扩展。但与代表词合并的词语必须出现在同一个聚类簇中。

4） 生成用户标签

我们按照一定的规则排序聚类后形成的簇，一种是按照簇内词语 TextRank 分数的加和，一种是按照簇内词语平均的 TextRank 分数。其对应的聚类代表词串，便是我们自动生成标签的顺序。

**Given**: a set of $words = \{w_1, ..., w_n\}$

**for** i =1 to n **do**

$C_i = \{w_i\}$

**end**

$C = \{C_1, ..., C_n\}$

j=1

**while** |C|>1

$(C_{n1}, C_{n2}) = \arg\min_{(C_u, C_v) \in C \times C} d(C_u, C_v)$

$C_j = C_{n1} \bigcup C_{n2}$

$C = C \setminus \{C_{n1}, C_{n2}\} \bigcup \{C_j\}$

j=j+1

图 4-5 词语层次聚类算法

## 4.3.3 实验与结果分析

### 4.3.3.1 实验数据与评价方法

测试用户依旧沿用 TextRank 方法中的 40 位用户。

评价指标仍是 P@N。评价指标有二：1）生成结果是否适合作为用户标签；2）生成结果是否反映了用户的兴趣。

评价方法也与 4.2 小节一致。

### 4.3.3.2 结果与分析

1） 词语间距离度量

我们借助爬取的百度百科页面进行词语语义相关度的计算，一共抽取了 1458050 条有效的百科词条信息。

我们对百科文章设置了一个大小为 4 的无序窗口，两个词语若同时出现在其中，则认定他们共现了一次。诸如“的、了”之类的停用词虽然对内容分析没有太大的帮助，但是，却有利于鉴别两个词语之间的语义关联度是否紧密。因此，我们仅对百科词条信息去除了标点符号，得到总词数 761352155 次，4475998 种词语。词语间的距离利用互信息如公式(4-5)计算。

表 4-3 是 40 位测试用户微博用词配对中互信息排序示例。

从排名前 20 的词对，我们可以看出，使用互信息作为词之间的距离是比较合适的。互信息排名靠前的分数都具有较强的关联度。

表 4-3 测试用户微博用词配对互信息排序示例

| 排序 | 词对 | 排序 | 词对 |
|---|---|---|---|
| 1 | 卡尔顿 富力丽 | 11 | 经济舱 头等舱 |
| 2 | 通途 天堑 | 12 | 李弘基 张根锡 |
| 3 | 付辛博 井柏然 | 13 | 汪东城 唐禹哲 |
| 4 | 宋楚瑜 亲民党 | 14 | 辰亦儒 汪东城 |
| 5 | 刘心 李炜 | 15 | 李炜 陈翔 |
| 6 | 兵库县 尼崎 | 16 | 刘亦菲 杨幂 |
| 7 | 评论员 杨禹 | 17 | 吴尊 辰亦儒 |
| 8 | 炎亚纶 辰亦儒 | 18 | 电锯 惊魂 |
| 9 | 奶糖 白兔 | 19 | 万湾 油菜花 |
| 10 | 李弘基 张根硕 | 20 | 朝天门 解放碑 |

2） 聚类终止条件

我们知道，层次聚类中簇间距离的算法有多种，例如单连接、全连接、均值算法等等。由于层次聚类时空复杂性本身就比较高，我们不考虑使用均值算法。单连接算法，使用簇内对象的最小距离作为簇间距离，当两簇之间出现中间点的时候极易发生合并，容易产生链状倾向。因此，我们的方法中使用全连接算法。

我们需要探讨一下聚类终止条件，即聚类过程达到什么条件，实际就可以停止了。我们假定，当簇间的距离小于某值时，我们便停止迭代。因此，我们需要寻找到这一阈值。我们将 40 位测试用户产生的词对的互信息分数由高到低进行排序，发现当互信息得分不超过 7 时，词对的语义关联已经较为微弱。因此，我们将阈值 K 粗略地设置为 7、6、5、4、3、2、1（它们为互信息分数）进行人工比较。通过观察发现，当阈值设置为 6 时，效果普遍较好。

我们将 K 值设置为 6。聚类效果如表 4-4 所示。

表 4-4 词语聚类示例

| 聚类簇序号 | 聚类簇内词语 |
|---|---|
| 簇 1 | 核心 价值观 高层 管理者 领导力 团队 人力资源 人力 企业 员工 绩效 |
| 簇 2 | 人才 毕业生 学生 学校 课程 学员 大学 专业 研究生 学历 |
| 簇 3 | 关键 领域 需求 客户 市场 产品 |
| 簇 4 | 奖金 福利 薪酬 工资 待遇 |

3） 标签自动生成效果

我们选取每个簇内 TextRank 分数最高的词语作为该簇的代表词，进行扩展。抽取的关键词，若按照簇内 TextRank 分数加和排序，我们称之为 cluster-sum；若按照簇内 TextRank 平均值排序，我们称之为 cluster-avg。同时，我们选取 baseline 系统与其对照。标签自动生成效果如表 4-5 所示。

表 4-5 基于聚类分析的标签自动生成效果(%)

a) 评价者 A 的评价结果

| | TFIDF | cluster-sum | cluster-avg |
|---|---|---|---|
| P@1 | 55.00 | **67.50** | 57.50 |
| P@2 | 48.75 | **56.25** | 51.25 |
| P@3 | 41.67 | **48.33** | 45.00 |
| P@4 | 40.63 | **49.38** | 41.88 |
| P@5 | 37.50 | **47.50** | 39.50 |
| P@6 | 35.42 | **45.42** | 39.17 |
| P@7 | 32.50 | **43.93** | 36.43 |
| P@8 | 29.69 | **41.88** | 35.94 |
| P@9 | 29.44 | **39.44** | 33.89 |
| P@10 | 28.25 | **38.75** | 32.25 |

b) 评价者 B 的评价结果

| | TFIDF | cluster-sum | cluster-avg |
|---|---|---|---|
| P@1 | 50.00 | **70.00** | 67.50 |
| P@2 | 43.75 | **60.00** | 56.25 |
| P@3 | 40.83 | **54.17** | 52.50 |
| P@4 | 38.13 | **53.75** | 50.63 |
| P@5 | 36.50 | **51.00** | 46.50 |
| P@6 | 36.25 | **48.75** | 45.83 |
| P@7 | 33.93 | **45.71** | 43.57 |
| P@8 | 31.88 | **44.06** | 41.56 |
| P@9 | 31.39 | **42.22** | 40.00 |
| P@10 | 29.75 | **41.00** | 38.00 |

两位评价者的评价一致性见表 4-6。

表 4-6 评价者 A、B 对标签生成结果评定的一致性

| | TFIDF | cluster-sum | cluster-avg |
|---|---|---|---|
| Cohen's kappa | 0.5770 | 0.5779 | 0.5480 |

从上述结果，我们可以看出，基于聚类分析的标签自动生成方法优于 Baseline，在 Top10 上提高了 10%左右。而按簇内权重加和的排序方式要优于按簇平均权重排序的方式。我们取 4.2 小节中 TextRank 分数排序前 200 的词语进行聚类，我们猜想，他们的分数差异并不是特别大，因此按平均权重排序未必能取得好的效果。

## 4.4 两种方法对比与分析

我们对基于 TextRank 和基于聚类分析自动生成用户标签的方法进行一下对比（选取每种方法中最优的方式）。他们的生成效果对照如表 4-7 所示。

表 4-7 基于 TextRank 与聚类分析生成标签效果对比

a) 评价者 A 的评价结果

| | **Baseline** | **TextRank** | **聚类分析** |
|---|---|---|---|
| P@1 | 55.00 | 62.50 | **67.50** |
| P@2 | 48.75 | **56.25** | **56.25** |
| P@3 | 41.67 | 47.50 | **48.33** |
| P@4 | 40.63 | **50.63** | 49.38 |
| P@5 | 37.50 | **47.50** | **47.50** |
| P@6 | 35.42 | **45.42** | **45.42** |
| P@7 | 32.50 | **44.29** | 43.93 |
| P@8 | 29.69 | **42.50** | 41.88 |
| P@9 | 29.44 | **40.28** | 39.44 |
| P@10 | 28.25 | 38.25 | **38.75** |

b) 评价者 B 的评价结果

| | **baseline** | **TextRank** | **聚类分析** |
|---|---|---|---|
| P@1 | 50.00 | 62.50 | **70.00** |
| P@2 | 43.75 | 52.50 | **60.00** |
| P@3 | 40.83 | 45.83 | **54.17** |
| P@4 | 38.13 | 48.13 | **53.75** |
| P@5 | 36.50 | 48.50 | **51.00** |
| P@6 | 36.25 | 47.50 | **48.75** |
| P@7 | 33.93 | 44.64 | **45.71** |
| P@8 | 31.88 | 42.50 | **44.06** |
| P@9 | 31.39 | 41.11 | **42.22** |
| P@10 | 29.75 | 40.25 | **41.00** |

就生成标签效果而言，两种方法都比我们的 baseline 系统（使用 TFIDF）的方法提高了将近 10 个百分点。而基于聚类分析的方法略优于基于 TextRank 的方法。我们提出基于聚类分析的方法，是为了避免同义标签堆积的现象。而通过观察测试用户生成的标签，我们发现确实解决了同义标签堆积的问题，使得生成的标签能在更多的维度上体现用户的兴趣。

我们猜想，基于聚类分析的方法相比 textrank 提高不多的原因在于，没有能够找到合适的中心词来代表聚类簇。我们仔细分析用于抽取标签的排名前 10 的聚类簇，查看该簇谈论的话题是否能够较准确地体现用户的兴趣。我们同样用 P@N 去衡量，结果如表 4-8 所示。结果证实了我们的猜想。

表 4-8 聚类簇反映用户兴趣的准确程度

| P@1 | P@2 | P@3 | P@4 | P@5 | P@6 | P@7 | P@8 | P@9 | P@10 |
|---|---|---|---|---|---|---|---|---|---|
| 75.00 | 76.25 | 72.50 | 71.88 | 69.50 | 65.42 | 62.86 | 61.56 | 60.28 | 58.75 |

## 4.5 本章小结

本章从关键词的角度为微博用户自动生成标签，介绍了两种方法：一种是基于 TextRank 的方法，一种是基于聚类分析的方法。我们使用真实的数据，对两种方法进行了实验，并分析了实验结果，同时进行了对比分析。

# 第5章 基于类别的标签自动生成

## 5.1 引言

第 4 章我们通过不同的关键词自动生成方法为微博用户添加标签。本章，我们将在类别这一粒度上自动生成用户标签。我们将尝试现有的自然语言处理方法，利用不同的资源，挖掘单层次及多层次类别的用户标签。

## 5.2 基于短文本分类的生成方法

本方法中，我们将借助于文本分类技术，预测出微博用户感兴趣的类别作为其标签。微博服务中，一般规定用户发布的微博文本长度不能超过 140 个字，这就注定了微博文本内容短小。而针对它的分类，属于短文本分类的范畴。

### 5.2.1 关键技术与原理

在这一部分，我们将简单介绍涉及本方法的一些关键技术和原理。

#### 5.2.1.1 文本分类

文本分类的任务可简单定义为：给定目标分类体系后，根据文本内容自动确定它的类别[49]。

文本分类作为一种典型的有监督机器学习问题，分为训练和分类两个阶段，如图 5-1 所示，具体过程如下[50]。

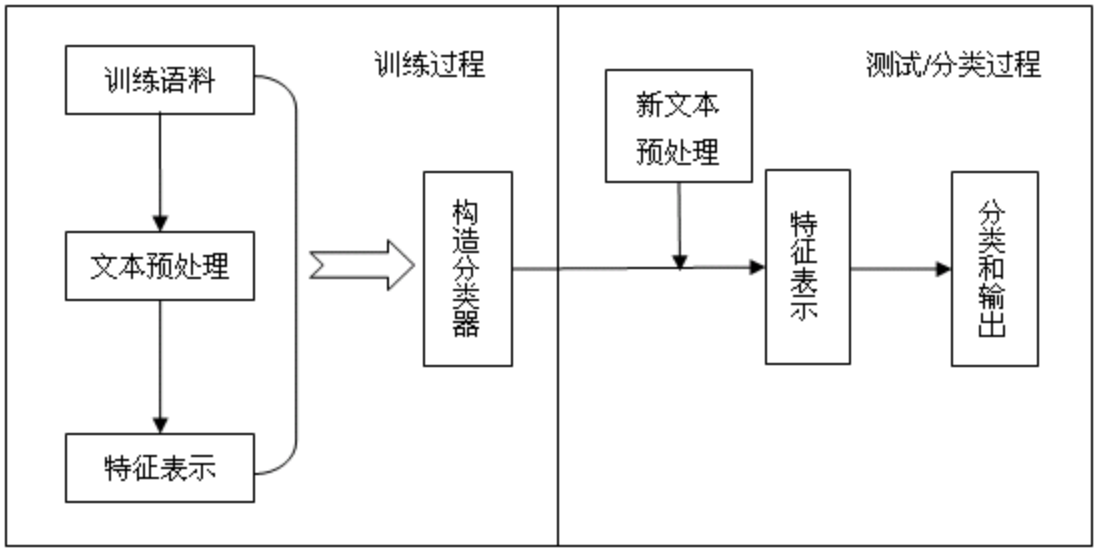



图 5-1 文本分类流程

**训练阶段：**

我们搜集一定规模的已经正确标注类别的文本集合作为训练语料。互联网网页，一般还需要抽取正文，得到纯文本的训练语料集合。

文本经过一定的预处理：一般对于英文文本来说，不需要分词但需要提取词干，对于中文文本，需要经过分词。之后再去停用词，例如各种没有意义的标点符号、虚词、副词、部分代词以及十分常用的词语等。通常我们会定义一个常用的停用词表。

预处理后，确定特征单位例如词语，用一定的描述模型如向量空间模型（Vector Space Model，VSM）[51]表示文本。选择合适的分类器，学习分类模型。

**分类阶段：**

未知类别的文本通过同样的预处理和特征表示，输入学习好的分类器模型中，将得到模型预测的类别。

文本分类主要涉及到以下几个关键问题。

1）文本表示

Salton 等人于 20 世纪 60 年代末提出了向量空间模型 VSM 的这一概念，也就是使用向量表示文本，该模型在 NLP 领域取得了广泛的应用，已成为最简便、高效的文本表示模型之一。

我们通常选用词语作为文本的特征项，多数使用 TFIDF 确定其权重，公式如（5-1）所示。

$$w(t,\bar{d})=\frac{tf(t,\bar{d})\times\log(N/n_t+0.01)}{\sqrt{\sum_{t=1}^{n}\left[tf(t,\bar{d})\times\log(N/n_t+0.01)\right]^2}} \tag{5-1}$$

其中，$tf(t,\bar{d})$ 为词项 t 在文本中的词频，$n_t$ 为训练语料中包含词项 t 的文本数目，N 为训练语料中文本的总数，然后经过归一化处理。

2）特征降维

文本分类经常遇到的一个问题就是特征维数过高和特征稀疏。高维度的特征集将耗费更长的训练时间，可能导致更差的分类效果。因此，我们需要运用特征降维技术。目前，特征降维有两种方式：特征选择和特征重构。

**特征选择**：特征选择遵循一定的准则从高维度的原始特征集合中选择最能反映类别的相关特征。主要思想是去除普遍存在和类别区分度不大的特征。文档频率（DF）、互信息（MI）、信息增益（IG）和卡方统计（CHI）等是主要的特征选择方法。

相关文献[52]表明，特征选择方法的有效性由高到低的排列顺序为：CHI、DF、MI、IG。而在中文环境中，文档频率与卡方统计的组合方法能获得更加的分类效果。在我们的方案中，将采用这样的组合方法。下面，仅简单介绍下 DF 和 CHI 的特征选择方法。

(1) 文档频率（DF）

若一个特征项文档频率过高，那么它极有可能在多个类别中都频繁出现，类别区分能力很差，可将其移除。

(2) CHI 统计

该统计方法用于衡量特征项 t 与文本类别 c 之间的相关度。值越大，说明相关度越高，公式如(5-2)所示。

$$\chi^2(t,c)=\frac{N\times(AD-CB)^2}{(A+C)(B+D)(A+B)+(C+D)} \tag{5-2}$$

其中，A 表示包含 t 且属于类 c 的文本数，B 表示包含 t 但不属于 c 的文本数，C 表示属于 c 但不包含 t 的文文本数，D 是既不属于 c 也不包含 t 的文本数。N 为训练语料的文本总数。

通常，一个特征项的卡方统计值取计算得到的类别中的最大的值。

**特征重构：**主要是指将特征空间进行变换，生成维度更小的特征空间，例如潜在语义标引 LSI、主成分分析等。

#### 5.2.1.2 短文本分类

微博文本、短信、论坛帖子、互联网查询等，一般都属于短文本。针对它们的分类，普通的文本分类技术基本都能解决。但是由于它们内容简短，通常不能提供足够的特征用于计算相似度，使得分类效果不够好。

很多研究工作专注于通过丰富短文本的语义特征，来提高分类效果。例如，有些学者利用伪相关反馈技术，借助搜索引擎返回的与短文本最相关的 N 条结果的 snippet 或者正文来扩充短文本。得到一个近似的长文本后，便可进行文本分类。还有些学者借助于电子知识库例如维基百科来丰富短文本的特征。

另外一种方法，是借助于主题模型等，来克服数据的稀疏性，在主题的层次上来进行文本分类。

#### 5.2.1.3 Latent Dirichlet Allocation

Latent Dirichlet Allocation（隐含狄利克雷分配，LDA 模型）由 Blei 等人于 2003 年提出的一种完全生成模型[53]。作为常用的隐含主题建模模型，它相对 LSI 和 PLSI 等具有清晰的层次结构。

LDA 模型是一种三层贝叶斯概率模型，词、主题和文档构成了它的三层结构。每篇文档被认为是 K 个隐含主题的混合分布，这是一个针对特定文档的多项式分布 $\Phi_d$；每个主题 z 被认为是一个词典 v 上的分布。文档的生成过程大致如下：

1） 对文本集合中的任一文档 d，生成文档长度 N， $N\sim Poisson(\varepsilon)$；

2） 对于文档 d，抽样得到 d 上 K 个隐含主题的多项式分布 $\Phi_d \sim Dir(\alpha)$

3） 考虑文档 d 中每一个词语 w：

采样生成 w 的主题， $z \sim Multi(\Phi_d)$

生成对应的词语， $w \sim Multi(\theta_z)$

LDA 的图模型如图 5-2 所表示。

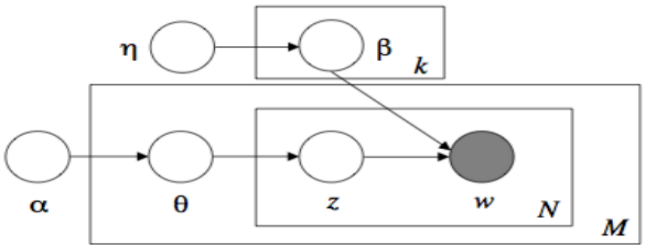



图 5-2 LDA 的图模型表示

#### 5.2.1.4 支持向量机（SVM）原理

本文使用支持向量机（Surport Vector Machine，SVM）算法训练文本分类模型。Vapnik 等人在 1995 年提出了 SVM，该算法在小样本、非线性和高维的分类问题上表现较为出色[54]。

统计学习理论的 VC 维理论以及结构风险最小原理是 SVM 的两大理论基础。利用有限的训练样本信息，以获得最好的泛化能力为目的，SVM 在模型的学习精度和学习能力之间寻找平衡。SVM 试图通过非线性变换算法即核函数将低维空间中线性不可分的样本映射到高维空间中使其线性可分。通俗来说，一个好的分类器，需要最大化分类间隔，如图 5-3 所示。SVM 就试图在特征空间中构建最优的分割超平面，使得分类模型在全局得到最优化。这就是 SVM 算法最主要的两个思想。

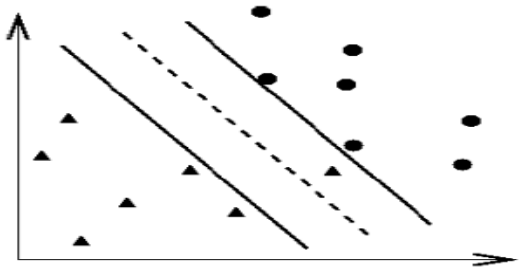


图 5-3 SVM 分类间隔

### 5.2.2 生成方法

我们将借助于短文本分类技术，为用户预测出感兴趣的类别作为标签。我们的方法整体框架如图 5-4 所示。

我们利用大规模的微博语料库，借助 LDA 对其进行主题分析。

在训练阶段，对微博训练语料使用之前得到的主题模型进行推理，得到主题层次上的特征，与词语层次上的特征相组合，构成特征空间，进行分类模型的训练。

在标签生成阶段，我们对用户发布的每一条微博，进行主题推理，之后抽取特征，用训练得到的分类器进行分类，得到相应的类别预测。用户的所有微博分类后，我们可以得到一个类别预测的集合。通过投票策略，可得知用户最感兴趣的类别作为其标签。

下面，我们将对主要步骤进行详细介绍。

1） 构建主题模型

这一步骤的关键是需要找到适合的语料进行主题模型构建。构建得到的模型，将用于之后的主题推理，从而构建主题层次上的特征。因此，我们需要较大规模的语料，从而保证它涵盖的主题面尽可能地广。另外，它的各种统计特征、语义特征等应该尽可能地适用于我们以后的应用场景。因此，我们并未选用大规模的网页库、知识库等等，而是使用微博语料来进行主题分析。

2） 确定分类体系，构建训练语料

据我们所知，目前，还没有公开的、具有一定规模的微博分类语料。因此，分类体系和训练语料都需要我们自己构建。

**构建类别体系的原则**是尽可能覆盖我们常见的微博文本类型，类别之间的覆盖度较低，能够较容易地寻找到相应类别的微博训练语料。

微博文本本身存在很多的噪声，对于普通的用户，他发布的微博倾向于涵盖很多类别。因此，我们将从比较正式的官方微博例如“新浪体育”、“新浪娱乐”或者人气很高的具有明显类别信息的微博例如“星座爱情”来抽取微博作为训练语料。

我们尽力符合之前提出的构建类别体系原则，将类别设置为十个，并搜集到一定规模和质量的微博训练语料。类别分别为：体育、娱乐、汽车、财经、时事/军事、科技、健康/养生、旅游/摄影/美食、星座/时尚/语录、校园/教育/职场。

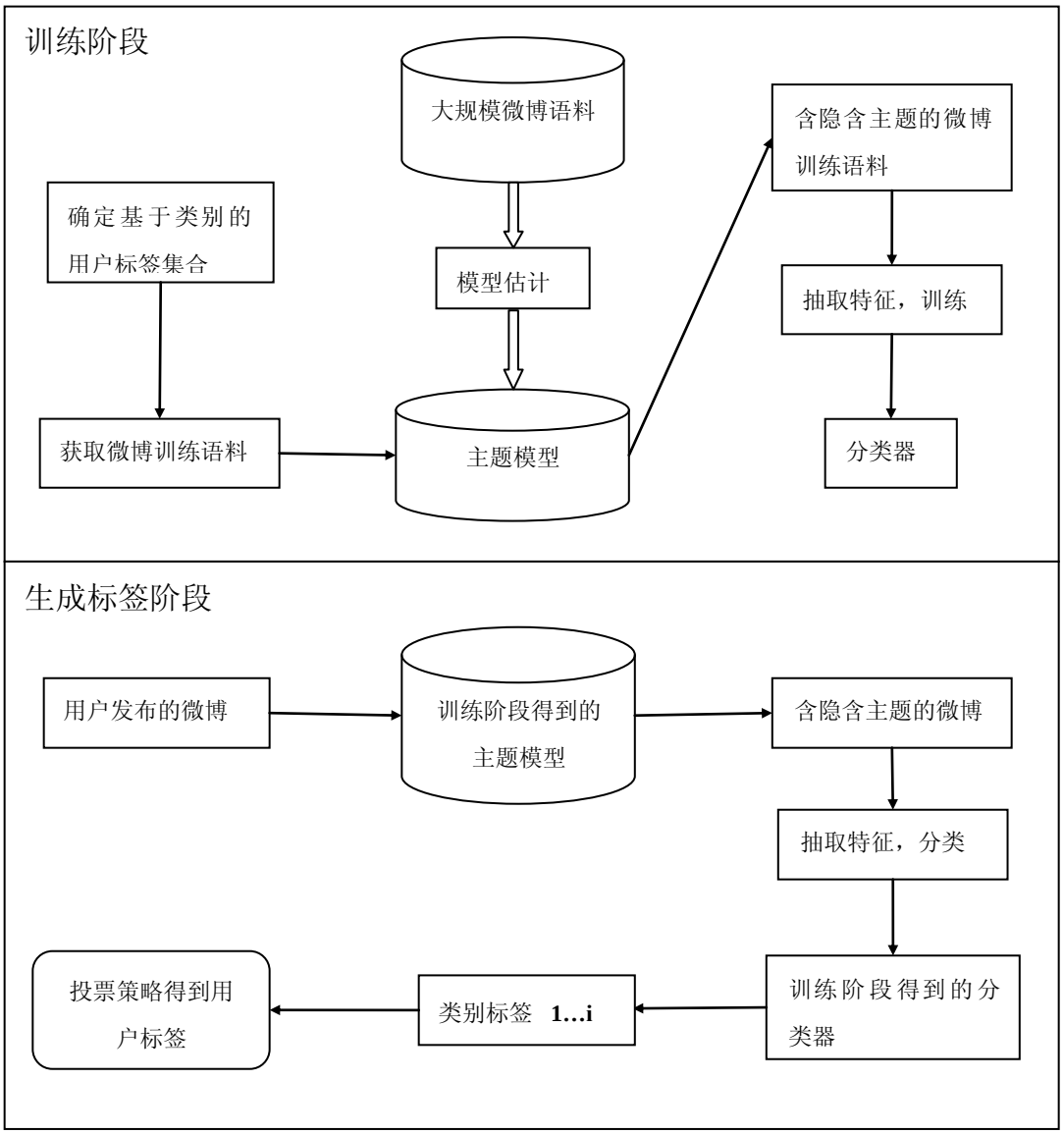



图 5-4 基于短文本分类的标签自动生成方法框架

3）构建特征空间

我们采取词语层次上的特征与主题层次上的特征相组合的策略，来构建训练分类器的特征向量空间。

词语层次上的特征，我们采取文档频率与卡方统计相结合的特征选择方式，抽取出类别区分度较明显的词语作为特征；主题层次方面，我们将使用 LDA 模型推理得到的 N 个 topic 作为特征。

4）用户标签的生成

用户标签的自动生成过程，可形式化地描述为：

用户标签集合 $C=\{c_1,...,c_{10}\}$。给定某个用户 u，抽取其发布的微博文本集

合 $W=\{w_1,...,w_n\}$，文本数目为 n。使用训练的短文本分类器进行预测，得到 n 条微博文本对应的预测类别列表 $L=\{l_1,...,l_n\}$，其中 $l_i \in C$。我们在预测类别列表上定义一个计数函数 $count(x,L)$，其中，$x \in C$，返回该标签在列表中出现的次数。类别标签按公式(5-3)计算分数排序。

$$\text{rank}(c)=count(c,L),c \in C \tag{5-3}$$

由高到低排序，取 Top3，便是该方法为用户生成的标签。

## 5.2.3 实验与结果分析

### 5.2.3.1 实验数据

我们为用户自动生成的标签集合或者说微博文本分类的目标类别集合为{体育、娱乐、汽车、财经、时事/军事、科技、健康/养生、旅游/摄影/美食、星座/时尚/语录、校园/教育/职场}。

用于 LDA 主题模型建模的较大规模微博语料一共包括 16610813 条微博，最晚的创建时间为 2011 年 12 月 15 日，最早的创建时间为 2007 年 2 月 2 日（该条微博可能是新浪微博服务测试时期发布的）。

上述语料是采取广度优先策略爬取，爬取到的每个用户的微博数目极少。1600 多万条微博，由 4514803 个用户发布，每位用户仅爬取到 3.68 条微博。这样也保证了该语料尽可能涵盖的话题广。

我们从新浪微博的某些官方微博和具有明显类别信息的人气较高的微博上抽取文本作为训练语料。训练和测试语料的微博文本数目如表 5-1 所示

表 5-1 训练和测试语料微博文本数目

| 类别标签 | 训练文本数 | 测试文本数 |
|---|---|---|
| 体育 | 45940 | 22963 |
| 娱乐 | 44496 | 22244 |
| 汽车 | 27802 | 13897 |
| 财经 | 43396 | 21702 |
| 时事/军事 | 29911 | 14952 |
| 科技 | 35817 | 17911 |
| 健康/养生 | 39529 | 19777 |
| 旅游/摄影/美食 | 22769 | 11378 |
| 星座/时尚/语录 | 57541 | 28757 |
| 校园/教育/职场 | 32919 | 16462 |

### 5.2.3.2 评价方法

1）分类效果

本文对短文本分类效果的评价采用信息检索领域常用的三个评价指标，准确率（precision）、召回率（recall）以及 F 测度。

假设测试语料中有属于类别 C 的 M 条微博，经过训练得到的分类器分类后，结果如表 5-2 所示。那么，针对于 C 类的分类效果评价指标：准确率、召回率、F 值分别如公式(5-4)、(5-5)、(5-6)。

表 5-2 分类状况示例

| | 属于 C 类数目 | 不属于 C 类数目 |
|---|---|---|
| 判定属于 C 类 | a | b |
| 判定不属于 C 类 | c | d |

$$precision = \frac{a}{a+b} \tag{5-4}$$

$$recall = \frac{a}{a+c} \tag{5-5}$$

$$F = \frac{2*precision*recall}{precision+recall} \tag{5-6}$$

同时，我们使用宏平均指标来度量分类器在所有类别（m 个类别）上的分类性能，准确率、召回率和 F 的宏平均值见公式(5-7)、(5-8)、(5-9)：

$$MacroP = \frac{\sum_{i=1}^{m} precision_i}{m} \tag{5-5}$$

$$MacroR = \frac{\sum_{i=1}^{m} recall_i}{m} \tag{5-6}$$

$$MacroF = \frac{\sum_{i=1}^{m} F_i}{m} \tag{5-7}$$

2）自动生成标签效果

在该自动生成标签的方法中，我们将为用户生成 3 个标签，反映其最感兴趣的类别。同样，通过人工标注，得到 P@N 指标。

测试用户依然沿用之前的 40 位微博测试用户。

#### 5.2.3.3 结果与分析

1）LDA 主题建模质量

我们使用 GibbsLDA++工具包对 1600 多万条微博进行建模，工具包将输出 4 个主要的文件：

<model_name>.phi：输出词语-主题分布。每行表示一个主题 topic，每列表示一个词，其对应位置输出的是相应的概率值。

<model_name>.theta：输出主题-文档分布。每行代表一篇文档，每一列代

表一个 topic，其对应位置输出的是相应的概率值。

<model_name>.tassign：每一行代表一个文档，每一列代表对应的词最有可能赋予的主题。

<model_name>.twords：输出每个主题中概率最大的前几个词语。

我们主要将借助于输出的<model_name>.twords 文件，评价主题建模质量。我们将主题数目经验地设置为 200 个。表 5-3 是部分主题-词示例。

表 5-3 主题-词示例

| 隐主题序号 | 隐主题生成概率最大的十个词 |
|---|---|
| Topic 1 | 人民 利比亚 卡扎菲 贺电 平民 军队 半岛 联合国 石油 政府 |
| Topic 25 | 旅游 酒店 旅行 温泉 享受 海南 门票 游泳 豪华 之旅 |
| Topic 28 | 明星 娱乐 签名 八卦 杨幂 偶像 代言 美人 演员 造型 |
| Topic 31 | 爱情 星座 狮子 天蝎 双鱼 射手 白羊 双子 爱上 处女 |
| Topic 59 | 汽车 开车 自行车 上海 宝马 二手车 车展 买车 奔驰 车主 |
| Topic 61 | 经济 银行 人民币 通胀 货币 加息 政策 金融 央行 市场 |
| Topic 66 | 大学 学生 学校 毕业 教育 教授 清华 学院 大学生 校长 |
| Topic 122 | 鸡蛋 食物 早餐 水果 土豆 牛奶 花生 蔬菜 豆腐 营养 |
| Topic 149 | 运动员 比赛 足球 黄健翔 巴萨 球迷 李娜 篮球 冠军 球员 |
| Topic 198 | 游戏 客户 软件 功能 iPhone 更新 同步 iTunes 手机 版本 |

同时，我们将根据每个主题的 Top-10 生成概率最大的词语，判断该主题是否有含义，我们进行人工评价。主题质量如表 5-4 所示。表中数据说明，LDA 建模的质量还是比较高的，生成的大部分主题是具有明显意义的。

表 5-4 主题建模质量

| 主题数目 | 有含义主题数目 | 无含义主题数目 | 有含义主题所占比例 |
|---|---|---|---|
| 200 | 165 | 35 | 82.5% |

3) 微博文本分类效果

我们将分别使用文档中词语最有可能被赋予的主题、主题-文档分布作为主题层次的特征与词语层次特征相结合，训练 SVM 分类器，得到分类器 word+topic-1、word+topic-2。

此外，仅使用词语特征构建的普通的文本分类器，我们称之为 wordLevel 分类器，使用文档中词语最有可能被赋予的主题作为特征的分类器我们称为 topic-1，使用主题-文档分布作为特征的分类器我们称之为 topic-2 分类器。

表 5-5、5-6、5-7、5-8、5-9 分别显示了相关的分类器的分类性能。

表 5-5 word+topic-1 分类效果

| 类别 | 准确率(%) | 召回率(%) | F 测度(%) |
|---|---|---|---|
| 体育 | 95.31 | 94.97 | 95.14 |
| 娱乐 | 88.07 | 88.49 | 88.28 |
| 汽车 | 89.97 | 85.50 | 87.68 |
| 财经 | 84.35 | 83.72 | 84.04 |
| 时事/军事 | 92.23 | 89.90 | 91.05 |
| 科技 | 85.38 | 85.36 | 85.37 |
| 健康/养生 | 87.88 | 84.22 | 86.01 |
| 旅游/摄影/美食 | 84.49 | 81.42 | 82.93 |
| 星座/时尚/语录 | 79.70 | 89.76 | 84.43 |
| 校园/教育/职场 | 84.86 | 78.79 | 81.71 |
| 宏平均 | 87.22 | 86.21 | 86.66 |

表 5-6 word+topic-2 分类效果

| 类别 | 准确率(%) | 召回率(%) | F 测度(%) |
|---|---|---|---|
| 体育 | 95.42 | 94.84 | 95.13 |
| 娱乐 | 87.98 | 88.44 | 88.21 |
| 汽车 | 89.29 | 85.78 | 87.50 |
| 财经 | 85.17 | 83.57 | 84.36 |
| 时事/军事 | 92.42 | 90.29 | 91.34 |
| 科技 | 85.86 | 84.98 | 85.42 |
| 健康/养生 | 87.56 | 83.81 | 85.64 |
| 旅游/摄影/美食 | 84.27 | 81.13 | 82.67 |
| 星座/时尚/语录 | 78.65 | 89.99 | 83.94 |
| 校园/教育/职场 | 84.97 | 77.95 | 81.31 |
| 宏平均 | 87.16 | 86.08 | 86.55 |

表 5-7 topic-1 分类效果

| 类别 | 准确率(%) | 召回率(%) | F 测度(%) |
|---|---|---|---|
| 体育 | 92.49 | 90.96 | 91.72 |
| 娱乐 | 78.52 | 81.05 | 79.76 |
| 汽车 | 87.35 | 79.59 | 83.29 |
| 财经 | 75.13 | 76.48 | 75.80 |
| 时事/军事 | 82.77 | 82.03 | 82.40 |
| 科技 | 76.33 | 75.58 | 75.95 |
| 健康/养生 | 82.92 | 77.87 | 80.31 |
| 旅游/摄影/美食 | 70.07 | 72.06 | 71.05 |
| 星座/时尚/语录 | 73.34 | 80.97 | 76.97 |
| 校园/教育/职场 | 73.30 | 67.58 | 70.32 |
| 宏平均 | 79.22 | 78.42 | 78.76 |

表 5-8 topic-2 分类效果

| 类别 | 准确率(%) | 召回率(%) | F 测度(%) |
|---|---|---|---|
| 体育 | 93.33 | 89.59 | 91.42 |
| 娱乐 | 78.32 | 78.23 | 78.27 |
| 汽车 | 84.44 | 80.61 | 82.48 |
| 财经 | 75.47 | 74.80 | 75.14 |
| 时事/军事 | 81.79 | 80.19 | 80.98 |
| 科技 | 80.61 | 70.64 | 75.29 |
| 健康/养生 | 85.92 | 74.33 | 79.71 |
| 旅游/摄影/美食 | 72.89 | 65.35 | 68.92 |
| 星座/时尚/语录 | 62.79 | 87.41 | 73.09 |
| 校园/教育/职场 | 76.00 | 61.48 | 67.97 |
| 宏平均 | 79.16 | 76.26 | 77.33 |

表 5-9 wordLevel 分类效果

| 类别 | 准确率(%) | 召回率(%) | F 测度(%) |
|---|---|---|---|
| 体育 | 95.75 | 94.78 | 95.26 |
| 娱乐 | 86.78 | 88.41 | 87.59 |
| 汽车 | 89.42 | 84.53 | 86.91 |
| 财经 | 85.13 | 85.03 | 85.08 |
| 时事/军事 | 91.92 | 89.61 | 90.75 |
| 科技 | 85.26 | 85.08 | 85.17 |
| 健康/养生 | 83.77 | 81.12 | 82.42 |
| 旅游/摄影/美食 | 77.59 | 75.45 | 76.51 |
| 星座/时尚/语录 | 74.76 | 83.78 | 79.01 |
| 校园/教育/职场 | 80.63 | 75.13 | 77.78 |
| 宏平均 | 85.10 | 84.29 | 84.65 |

从上述表格中，我们可以看到，使用普通的文本分类技术，分类效果已经不错，表现并不像我们想象中的那么差。分析原因，可能是因为我们搜集的微博训练语料规模已经较大，使得训练的分类器能够较好地捕捉到待分类的短文本中的特征。

单独使用主题作为特征的分类器，利用文档中最有可能赋予词语的主题作为特征，效果略高于使用主题-文档分布。但是，两者的表现都不如 wordLevel 分类器。直觉上，由于短文本特征较稀疏，我们通过 LDA 降维后，可能会收到更好的效果。分析原因，可能是因为我们设置的主题数目不大合适。

由于用于 LDA 建模的微博语料数量已经上百万，整个建模过程非常缓慢，因此本文并没有通过实验验证主题数目怎样设置比较合理。而是通过主题和词语特征相结合，构造分类器来实现短文本分类。我们可以看到，本方法中的训练器结果都优于之前讨论过的分类器，提高了将近 2 个百分点。同样地，使用主题-文档分布作为特征，结果稍差。

这几个分类器在“旅游/摄影/美食”、“星座/时尚/语录”和“校园/教育/职

场”三个类别上表现都稍差，原因大致是这个三个类别中的每一类其实涵盖的话题都比较多比较杂。这是由选取的微博语料特征决定的。在构建训练语料时，我们已经尽力使得每个类别倾向于正交，但是，通过对带有类别信息的微博账号发布的文本的仔细分析，我们认为这三个类别的每一类已经无法再细分。

虽然我们选取的构建微博训练语料的微博用户发布的所有微博已经倾向于属于同一个类别，但是，不可避免地，会引入一些噪声。每个用户发布的微博都具有自己的风格。而微博测试语料也是源自于他们。为了测试训练的分类器的泛化能力，从测试用户的中随机抽取了 200 条，使用效果最好的分类器进行分类，人工标注分类是否正确，分类准确率达到 80.5%。因此，我们认为，分类器具有一定的分类微博的能力。我们将借助其进行用户标签自动生成。

3） 标签生成效果

我们使用训练得到的分类器对测试用户的微博进行分类，通过我们的投票策略，输出最感兴趣的三个类别作为其标签，人工判定是否准确。由于普通微博用户的原创微博时常描述一些日常琐事，类别信息可能不显著。因此，我们使用了两种策略，一是对用户发布的所有微博进行分类；二是对用户转发的微博进行分类，结果如表 5-10 所示。同样，我们有两位评价者对生成结果进行判定。

表 5-10 基于短文本分类的标签生成效果(%)

a) 评价者 A 的评价结果

| | P@1 | P@2 | P@3 |
|---|---|---|---|
| 使用所有微博 | 100.00 | 90.00 | 78.33 |
| 使用转发微博 | 97.50 | 92.50 | 79.17 |

b) 评价者 B 的评价结果

| | P@1 | P@2 | P@3 |
|---|---|---|---|
| 使用所有微博 | 97.50 | 80.00 | 70.00 |
| 使用转发微博 | 100.00 | 90.00 | 72.50 |

表 5-11 是两位评价者评定结果的一致性。

表 5-11 评价者 A、B 对标签生成结果评价的一致性

| | 使用所有微博 | 使用转发微博 |
|---|---|---|
| Cohen's kappa | 0.4124 | 0.4828 |

两位评价者对结果的评定一致性依旧处在 0.40 至 0.75 的范围内，比较不错。但是，低于第 4 章我们对基于词语的标签生成结果一致性的判定。我们猜测原因，候选标签集合中标签数目虽然少，但是人工评价出用户最感兴趣的三个类别还是具有很大的主观性。

从表 5-10 的结果看，基于短文本分类的标签生成效果较好。而且，使用转发微博略优于使用所有微博，再一次验证了转发微博更能体现用户兴趣。

但是，我们基于短文本分类的方法也有自身的弱点。分类体系是自己设置的，可能出现设置不够合理的状况。同时，分类的粒度稍粗，导致用户的候选标签集合不够大。

## 5.3 基于百度百科的生成方法

### 5.3.1 百度百科介绍

百度百科是由百度公司在 2008 年 4 月 21 日正式发布的一部中文知识性百科全书。百度百科是一部自由、开放的网络百科全书，它提供的所有服务，例如搜索、阅读、创建和编辑词条等等，互联网用户都可以免费使用。也可以说，它是一部由网民共同协作编写的百科全书。

为了保证词条的质量，百度公司规定只有百科注册用户才能参与词条、评论的创建、编写。同时，用户的所有操作提交后，必须通过内部管理员的人工审核，操作才会有效，相应改动的内容才能被公开。为了减少百科内容被恶意编辑、篡改的事件发生，百度对用户的编辑权限也做出了一定的限制：不同等级的用户拥有的操作权限不同。另外，百科还引入了权威认证的机制，对词条内容进行专业认证，保证其权威性。通过以上举措，我们目前阅读到的百度百科可以说是一部质量较高的中文百科全书。而它的词条数目也早已经突破了 200 万。

构成百度百科的基础内容是词条。一个词条页面大致可以分为百科名片、词条正文、开放分类、相关词条、参考资料和扩展阅读这个几个部分，根据每个词条的具体情况，某些部分可以没有。其中，百科名片是词条的概括性描述；词条正文是可以由多个段落组成的对词条的详细描述；开放分类提供词条的属性标签，最多有 5 个，通常这种标签具有一定的类别信息；相关词条是与当前词条联系比较紧密的相关条目，而一般这种联系必须是横向的，例如“乔峰”的相关词条可以是“段誉”、“阿朱”、“虚竹”等，但不可以是“天龙八部”。图 5-5 是开放分类和相关词条的示例。


图 5-5 百度百科词条页面开放分类和相关词条示例

尽管在百科的词条页面上，我们只能看到该词条的开放分类标签，但实际上，在百科内部，有一个具有层次的三层分类体系。第一层具有十二个大类；每个大类下面又包含若干个中类，这就是第二层分类；每个中类下面又可细分为若干个小类。图 5-6 显示的是分类体系的十二个大类。


图 5-6 百度百科十二个大类

## 5.3.2 生成方法

我们将百度百科三层分类体系第三层的小类别作为查询，在百度百科中将搜索得到该类别下的词条页面。但是，通过这种方式，对于每一个小类别，我们最多只能获取 760 个相关的词条页面，而更多的实际上可以被划分为该类别的词条页面，我们获取不到。如图 5-7 所示，我们搜索“技术_互联网_电子商务”类别下的页面，结果显示共有词条 3469 个，但是，我们遍历搜索结果页面，最多只能查看到 760 个词条页面。在该方法中，我们需要使用百度百科的三层分类体系例如“技术_互联网_编程”、“生活 _娱乐_影视”作为用户的标签候选集合。因此，我们可利用的资源便是能够获取到的每个小类别下的最多 760 个词条页面。

我们通过分析用户发布的微博，将其最感兴趣的百科三层类别作为他的标签。方法的主要思想是：将用户发布的每一条微博映射到最相关的百科词条页面，获取其类别，通过一定的投票策略决策出用户的标签。


图 5-7 类别标签的搜索页面示例

下面，具体介绍一下方法中涉及的关键步骤。

1） 百科页面获取与处理

我们按照百度百科提供的开放分类浏览页面，爬取到所有第三层小类别的查询 URL，形如 http://baike.baidu.com/taglist?tag=****，该 URL 指向的就是

如图 5-7 类似的该类别标签的搜索页面。我们通过解析该搜索页面，下载得到百科开放的 760 个相关的词条页面。

获取页面后，并且按照第三层小类别分类存储后，我们进行正文的提取。词条页面一共有两种，一种是歧义页面，即一个词条包含多种义项，如图 5-8 所示；一种是非歧义页面。我们需要分别对其做处理：对于歧义页面，将每一个义项单独作为一个页面提取出来。

我们提取百科页面的标题、百科名片、正文、类别，进行分词等处理后，建立索引，待后续使用。


图 5-8 歧义页面示例

2） 搜索与微博相关的词条页面

给定一条微博，我们需要找到与其相似或者说相关的百科页面，以获取其类别。我们将这个问题转换成为搜索问题。从微博中提炼出查询，从索引过的百科页面中搜索出最相关的。

我们使用 Indri 对词条页面建立索引。Indri 源自 Lemur 系统，是 CMU 和 UMass 联合推出的一个用于语言模型和信息检索研究的系统。在这之上可以实现基于语言模型和传统的向量空间模型等的检索。Indri 早已经受到了学术界的广泛欢迎。

我们将分别抽取微博中的名词、名词及形容词的组合构成带有权重的查询，其权重采用 TFIDF 计算得到，构造的查询形如“#weight(0.45 巨蟹座 0.35 性格 0.20 特点)”。用构建的查询，我们搜寻到与该查询最相关的词条页面，并获取其对应的类别。

3） 用户标签生成

假定有 m 种具有三层分类的类别标签，我们将其作为用户的候选标签集合，表示为 $C=\{c_1,...c_i...,c_m\}$。

给定某个用户 u，抽取其发布的微博文本集合 $W=\{w_1,...,w_n\}$，文本数目为 n。我们分别对其构造查询，得到查询集合 $Q=\{q_1,...q_i\ ,..q_n$ 。对每一个查询

q，我们获取 Top N 个结果，作为该查询相关的词条页面。因为每个词条可能属于若干个类别，因此我们得到的与该查询（或者说其对应的微博）关联的类别可能不止 N 个。对关联的类别中的每一个类别 c，我们计算其分数，如公式(5-8)。其中， $freq(c)$ 指 c 在该次查询中出现的频数，而 d 是我们设置的一个常数，例如，我们将其设置为 1，则 c 的分数就是在该次查询中出现的次数；若将 d 设置为该类别关联的词条页面出现在 Top N 结果中的具体位置，则说明 c 的分数受到搜索结果排序的影响，排序越靠前的，对分数的贡献越大。

$$score(c) = \sum_{freq(c)} \frac{1}{d} \tag{5-8}$$

整个查询过程停止后，我们将对候选标签集合中的每一个标签 c 计算一个排序分数，见公式(5-9)，其中 n 就是指查询的次数。我们取排序前 10 的作为为用户自动生成的标签。

$$rank(c) = \sum_{n} score(c) \tag{5-9}$$

## 5.3.3 实验与结果分析

### 5.3.3.1 实验数据和评价方法

由于人物这一大类中包含太多的歧义页面，例如一个人名可能对应不同领域的多个人甚至同一领域的多个人，因此，在本方法中，略去该类。表 5-12 显示的三层类别体系中每一层的类别数目，第三层类别将作为用户的候选标签集合。表 5-13 显示的是包含页面最多和最少的 10 个类别。当一个类别下实际的词条数目超过 760 时，我们能够爬取到的也只是 760 个页面。我们一共索引了 342163 个有效的百科页面。

表 5-12 三层类别体系中每一层的类别数目

| 层数 | 类别数 |
|---|---|
| 第一层 | 11 |
| 第二层 | 88 |
| 第三层 | 1085 |

表 5-13 包含词条页面最多和最少的类别示例

| 类别标签 | 包含词条数目 | 类别标签 | 包含词条数目 |
|---|---|---|---|
| 科学_生物学_植物学 | 760 | 体育_射击_射击运动 | 1 |
| 地理_区域地理_内蒙古 | 760 | 科学_应用科学_区域社会学 | 1 |
| 自然_动物_昆虫 | 760 | 自然_微生物_螺旋体 | 1 |
| 自然_动物_鱼类 | 760 | 科学_生物学_悉生生物学 | 1 |
| 地理_区域地理_西藏 | 760 | 科学_生物学_生物数学 | 1 |
| 地理_地质_地质 | 760 | 体育_健美健身_搏击健美操 | 1 |
| 生活_娱乐_影视 | 760 | 自然_人类_蓝田人 | 1 |
| 文化_语种_语言 | 760 | 地理_水域_流水域 | 1 |
| 生活_旅游_名胜古迹 | 760 | 自然_人类_山顶洞人 | 1 |
| 生活_饮食_烹饪 | 760 | 体育_对抗运动_职业摔跤 | 1 |

同样，我们沿用之前的 40 位测试用户。为每一个人生成 10 个标签，人工进行评价，并计算出 P@N 指标。

#### 5.3.3.2 结果与分析

我们一共使用四种方式来生成用户标签：

Nouns+Adj：抽取微博中名词和形容词构造查询，排序公式(5-8)中 d 设置为 1；

Nouns+Adj+Weight：抽取微博中名词和形容词构造查询，d 设置为当前页面在搜索结果中的排序；

Nouns：抽取微博中名词构造查询，d 设置为 1；

Nouns+Weight：抽取微博中名词构造查询，d 设置为当前页面在搜索结果中的排序。

对于每条微博构造的查询，我们选取与其最相关的 3 个页面即搜索结果 Top3 的词条页面。上述四个方式生成用户标签的表现如表 5-14 所示。

表 5-14 基于百度百科用户标签生成效果(%)

a) 评价者 A 的评价结果

| | Nouns+Adj | Nouns+Adj+Weight | Nouns | Nouns+Weight |
|---|---|---|---|---|
| P@1 | **82.50** | **82.50** | 80.00 | 80.00 |
| P@2 | 75.00 | 76.25 | **78.75** | **78.75** |
| P@3 | 73.33 | 73.33 | 73.33 | **75.00** |
| P@4 | 72.50 | 72.50 | 72.50 | **73.13** |
| P@5 | 71.50 | 72.00 | **73.50** | 71.50 |
| P@6 | **72.08** | **72.08** | 71.25 | 70.42 |
| P@7 | 70.00 | 70.00 | 70.00 | **70.36** |
| P@8 | 68.75 | 69.38 | **70.00** | **70.00** |
| P@9 | 67.22 | 68.61 | **68.89** | 68.61 |
| P@10 | 67.00 | **67.75** | 67.25 | **67.75** |

b) 评价者 B 的评价结果

| | Nouns+Adj | Nouns+Adj+Weight | Nouns | Nouns+Weight |
|---|---|---|---|---|
| P@1 | **80.00** | **80.00** | **80.00** | **80.00** |
| P@2 | 71.25 | 70.00 | **73.75** | 72.50 |
| P@3 | 67.50 | 67.50 | **68.33** | **68.33** |
| P@4 | 65.63 | 65.63 | 65.63 | **66.88** |
| P@5 | 63.00 | 63.00 | 65.00 | **65.50** |
| P@6 | 63.33 | 62.92 | **65.00** | 64.58 |
| P@7 | 62.50 | 62.50 | **63.57** | **63.57** |
| P@8 | 61.88 | 61.25 | **64.69** | 64.38 |
| P@9 | 60.56 | 60.56 | 63.06 | **63.33** |
| P@10 | 60.25 | 61.00 | 61.50 | **62.00** |

表 5-15 两位评价者对结果评定的一致性度量。

表 5-15 评价者 A、B 对生成标签结果评定的一致性

| | Nouns+Adj | Nouns+Adj+Weight | Nouns | Nouns+Weight |
|---|---|---|---|---|
| Cohen's kappa | 0.5677 | 0.5686 | 0.5802 | 0.5808 |

上表中，四种方式的表现差不多。就抽取微博词语构造查询的方式而言，名词在整体上略优于名词与形容词的组合，或许是因为形容词的加入引入了部分的噪声。

就排序方式而言，对每一个类别施加与其搜索位置相关的权重，能够得到稍好的生成效果，但是，在人工评价中，我们发现，排序公式中权重的引入几乎只是影响了前 10 个类别标签的排序位置，而前 10 个类别标签集合中成员几乎不变。我们猜测，倘若权重设置得更加合适，能够带来更好的生成效果。

## 5.4 两种方法对比与分析

两种基于类别的方法都能取得不错的效果。仔细分析，其实这两种方法在效果上从指标上没有太大的可比性。基于短文本分类的方法，我们仅设置 10 个大类别作为用户的候选标签，而基于百度百科的方法，用户潜在的标签有 1085 种。候选标签集合的基数相差悬殊，因此从 P@N 指标上，我们不能明确地对比出谁好谁坏。

从生成标签的粒度而言，我们认为基于百度百科的三层类别标签能更好更细地识别出用户的兴趣。

## 5.5 本章小结

本章从类别的角度为微博用户自动生成标签，介绍了两种方法：一种是基于短文本分类的方法，一种是利于百度百科资源的生成方法。我们使用真实的数据，对两种方法进行了实验，并分析了实验结果，同时进行了对比分析。

# 结　论

近年来，微博作为新型的互联网应用，在国内外都受到了广大网民越来越多的关注。微博的广泛使用，也引起了学术界极大的兴趣。在自然语言处理、信息检索和社会计算领域，专注于微博的研究课题也逐渐开展并积累起来，例如微博社交网络的分析与社交圈的识别、微博事件和话题的检测、微博搜索、微博情感分析等等，但是，相比起来，用户标签自动生成相关的研究工作却刚刚开展。而研究微博用户标签及其自动生成，对深入了解用户兴趣、为其提供个性化服务以及针对微博用户的搜索等有着重大的意义。

本文着力于基于微博内容的用户标签自动生成：使用微博用户自己产生的文本内容（UGC），借助于微博文本分析，尝试从基于关键词和基于类别两个角度，为用户自动生成标签。我们试图从研究过程中更深入地了解用户兴趣，为用户自动生成能体现其兴趣的标签。

首先，为了更深入地了解用户为自己添加标签的行为，更好地掌握标签相关数据的性质、规律，本文对标签相关的数据做出了较为详细的分析。通过对较大规模的数据统计、分析，我们发现：1）用户在微博服务中使用标签功能的百分比较低，当一个用户发布、评论微博等行为较活跃的时候，他也倾向于会为自己添加标签。2）用户的标签集合具有长尾分布的规律，少数标签被使用的次数相当频繁，而绝大多数标签仅被极少数人使用过。这说明微博用户标签十分个性化，同时也意味着比较难组织、规范。3）用户标签绝大多数由单个的词语，或者长度较短的短语构成。使用频繁的标签往往带有比较明显的类别信息。同时，本文提出了基于文本的标签源这一概念，即可以用于生成用户标签的微博文本：用户的原创、转发、评论和收藏微博。通过具体的数据和实验，我们发现，不同标签源谈论的话题有所不同。转发、收藏的微博更能体现用户兴趣，而评论则较差。

其次，本文尝试从关键词的角度来为用户自动生成标签。本文介绍了基于 TextRank 和聚类分析的自动生成方法。基于 TextRank 的方法借助于微博文本中词语的共现关系，抽取出较为重要的词语，用于生成标签；基于聚类分析的方法，试图从多个兴趣维度挖掘出较为重要的词语，用于标签生成。实验结果表明，两者的生成效果都优于我们的 baseline（TFIDF）策略。

再次，本文尝试从类别的角度来为用户自动生成标签。本文介绍了基于短文本分类和百度百科的自动生成方法。基于短文本分类的方法，人工构建了微博训练语料，将 10 个目标类别作为用户标签的候选集合，选取用户感兴趣的类别作为其标签；基于百度百科的方法，使用百度百科的 1085 个三层分类类别作为用户标签的候选集合，利用词条信息，选取能体现用户兴趣的细粒度类

别作为用户的标签。实验结果表明，两者都能较好地体现用户的兴趣。

本文在基于内容分析的中文微博用户标签自动生成上做了一次尝试。当然，还存在其局限性和有待改进的地方。后续，我们一是需要改进基于内容分析的标签自动生成方法，二是我们可以通过获取大规模的数据，借助于社交网络分析，使研究工作更加完善。

# 参考文献

[1] 魏艳. Twitter 用户预计下月突破 5 亿年底达到 9 亿[EB/OL]. (2012-01-18) [2012-05-22]. http://it.people.com.cn/GB/16905781.html.

[2] 马晓宁. 中国微博客价值与发展研究[D]. 南昌大学硕士学位论文，2010：20-24.

[3] R. Jäschke, L. Marinho, A. Hotho, et al. Tag Recommendations in Social Bookmarking Systems[J]. AI Communications, 2008, 21(4): 231-247.

[4] Klaas Dellshaft, Steffen Staab. An Epistemic Dynamic Model for Tagging Systems[C]. In HT'08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia. 2008: 71-78.

[5] 毛军. 元数据、自由分类法（Folksonomy）和大众的因特网[J]. 现代图书情报技术，2006，133：1-5.

[6] Thomas Vander Wal. Explaining and Showing Broad and Narrow Folksonomies [EB/OL]. http://www.personalinfocloud.com.

[7] A. Hotho, R. Jäschke, C. Schmitz and et al. Information Retrieval in Folksonomies[C]: Search and Ranking, in: The Semantic Web: Research and Applications. 2006: 411–426.

[8] Manish Gupta, Rui Li, Zhijun Yin, et al. Survey on Socail Tagging Tagging Techniques[C]. Sigkdd Explorations. 2010, 12(1): 58-72.

[9] Morgan Ames, Mor Naaman. Why We Tag: Motivations for Annotation in Mobile and Online Media[C]. In Conference on Human Factors in Computing Systems, CHI. 2007: 971-980.

[10] Christopher H. Improved Annotation of the Blogosphere via Autotagging and Hierarchical Clustering[C]. WWW. 2006: 625-632.

[11] Marieke Guy, Emma Tonkin. Folksonomies: Tidying up Tags?[J]. D-Lib Maggzine. 2006, 12(1): 2-4.

[12] Simo Overell, Borkur Sigurbjornsson, Roelof van Zwol. Classifying Tags Using Open Content Resources[C]. WSDM. 2009: 64-73.

[13] Heymann P, Garcia Molinay H. Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems[R]. Technical Report InfoLab. Department of Computer Science, Stanford University, Stanford, CA, USA. April 2006: 1-5.

[14] 吴思竹. 社会标注系统中标签推荐方法研究进展[J]. 图书馆杂志. 2010，29(3)：48-52.

[15] 刘知远. 基于文档主题结构的关键词抽取方法研究[D]. 清华大学博士学

位论文. 2011：6-7.

[16] 靳延安. 社会标签推荐技术与方法研究[D]. 华中科技大学博士学位论文. 2011：12-15.

[17] Jaschke R, Marinho L, Hotho A, et al. Tag Recommendations in Folksonomies[C]. Proceedings of ECML/PKDD. 2007: 506–514.

[18] Haewoon Kwak, Changhyun Lee, Hosung Park, et al. What is Twitter, a Social Network or a News Media?[C]. WWW 2010. 591-600.

[19] S. Asur, B. A. Huberman, G. Szabo, et al. What Trends in Social Media - Persistence and Decay[C]. In 5th International AAAI Conference on Weblogs and Social Media. 2011: 434-437.

[20] M. Cha, H. Haddadi, F. Benevenuto, et al. Measuring User Influence in Twitter: The Million Follower Fallacy[C]. In Fourth International AAAI Conference on Weblogs and Social Media. 2010: 10-17.

[21] Takeshi Sakaki, Makoto Okazaki, Yutaka Matsuo. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors[C]. WWW 2010. 851-860.

[22] Kristina Lerman, Rumi Ghosh. Information Contagion: an Empirical Study of Spread of News on Digg and Twitter Social Networks[C]. AAAI. 2010: 90-97.

[23] Mario Cataldi, Luigi Di Caro, Claudio Schifanella. Emerging Topic Detection on Twitter based on Temporal and Social Terms Evaluation[C]. MDMKDD '10 Proceedings of the Tenth International Workshop on Multimedia Data Mining. 2010: 1-10.

[24] Michael Mathioudakis, Nick Koudas. TwitterMonitor: Trend Detection over the Twitter Stream[C]. International Conference on Management of Data - SIGMOD. 2010: 1155-1158.

[25] Saša Petrovic, Miles Osborne, Victor Lavrenko. Stream ing First Story Detection with Application to Twitter[C]. HLT '10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2010:181-189.

[26] Luciano Barbosa, Junlan Feng. Robust Sentiment Detection on Twitter from Biased and Noisy Data[C]. COLING.. 2010:36-44.

[27] Adam Bermingham, Alan F. Smeaton. Classifying Sentiment in Microblogs: is Brevity an Advantage?[C]. International Conference on Information and Knowledge Management - CIKM. 2010: 1833-1836.

[28] Jianshu Weng, Ee-Peng Lim, Jing Jiang, et al. TwitterRank: Finding Topic-Sensitive Influential Twitterers. Web Search and Data Mining - WSDM. 2010: 261-270.

[29] John Hanno, Mike Bennett, Barry Smyth. Recommending Twitter Users to

Follow Using Content and Collaborative Filtering Approaches[C]. RecSys. 2010: 199-206.

[30] Jilin Chen, Rowan Nairn, Les Nelson, et al. Short and tweet: Experiments on Recommending Content from Information Streams[C]. In: Proceedings of the 28th international conference on Human factors in computing systems (CHI '10), New York, NY, USA, ACM (2010) 1185–1194.

[31] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, et al. Comparing Twitter and Traditional Media Using Topic Models[C]. European Colloquium on IR Research - ECIR. 2011:338-349.

[32] Hong, L., Davison, B.D.: Empirical Study of Topic Modeling in Twitter. In: Pro-ceedings of the SIGKDD Workshop on SMA. 2010: 80-88.

[33] Daniel Ramage, Susan T. Dumais, Daniel J. Liebling. Characterizing Microblogs with Topic Models[C]. ICWSM. 2010: 130-137.

[34] Fabian Abel, Qi Gao, Geert-Jan. Sematic Enrichment of Twitter Posts for User Profile Construction on the Social Web[C]. ESWC. 2011: 1-15.

[35] Fabian Abel, Qi Gao, Geert-Jan. Analyzing User Modeling on Twitter For Personalized News Recommendations[C]. UMAP. 2011: 1-12.

[36] Fabian Abel, Qi Gao, Geert-Jan. TUMS: Twitter-based User Modeling Service[C]. ESWC. 2011: 1-15.

[37] Matthew Michelson, Sofus A. Macskassy. Discovering Users' Topics of Interest on Twitter: a First Look[C]. AND '10 Proceedings of the fourth workshop on Analytics for noisy unstructured text data. 2010:73-80.

[38] Yegin Genc, Yasuaki Sakamoto, Jeffrey V. Nickerson. Discovering Context: Classifying Tweets through a Semantic Transform based on Wikipedia[C]. HCII. 2011: 484-492.

[39] Wei Wu, Bin Zhang, Mari Ostendorf. Automatic Generation of Personalized Annotation Tags for Twitter Users[C]. ACL. 2010: 689-692.

[40] Theodoros Lappas, Kunal Punera, Tamas Sarlos.Mining Tags Using Social Endorsement Networks[C]. SIGIR. 2011:195-204.

[41] Yuto Yamaguchi, Toshiyuki Amagasa, Hiroyuki Kitagawa. Tag-based User Topic Discovery Using Twitter Lists[C]. Advances in Social Network Analysis and Mining - ASONAM. 2011:13-20.

[42] Balachander Krishnamurthy, Phillipa Gill, Martin Arlitt. a Few Chirps about Twitter[C]. Proceedings of the first workshop on online social networks. 2008:19-24.

[43] Rada Mihalcea, Paul Tarau. TextRank: Bringing Order into Texts[C]. EMNLP. 2004: 404-411.

[44] Lawrence Page, Sergey Brin, Rajeev Motwani, et al. The PageRank Citation Ranking: Bringing Order to the Web[C]. Standford. 1998: 1-17.

[45] Hulth A. Improved automatic keyword extraction given more linguistic knowledge[C]. EMNLP. 2003: 216–223.

[46] Daniel Jurafsky, James H. Martin.Speech and Language Processing.2008: 627-642.

[47] Evgeniy Gabrilovich, Shaul Markovitch. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis[C]. International Joint Conference on Artificial Intelligence - IJCAI. 2007:1606-1611.

[48] Jiawei Han, Micheline Kamber. 数据挖掘概念与技术. 机械工业出版社，2007：206-209.

[49] 刘挺，秦兵，张宇等. 信息检索系统导论. 机械工业出版社，2008：186-190.

[50] 庞剑锋，卜东波，白硕. 基于向量空间模型的文本自动分类系统的研究与实现. 计算机应用研究. 2001，18(9)：23-26.

[51] David D Lewis. Representation and learning in information retrieval [D]. Univ. of Massachusetts. 1992.

[52] 代六玲，黄河燕，陈肇雄. 中文文本分类中特征抽取的比较研究. 中文信息学报. 2004，18(1)：26-32.

[53] DM Blei, AY Ng. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research. 2003: 993-1022.

[54] Johan A. K. Suykens, Joos Vandewalle. Least Squares Support Vector Machine Classifiers[J]. Neural Processing Letters. 1999. 9(3): 293-300.

# 攻读学位期间发表的学术论文

# 哈尔滨工业大学学位论文原创性声明及使用授权说明

## 学位论文原创性声明

本人郑重声明：此处所提交的学位论文《面向微博用户的标签自动生成技术研究》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果。据本人所知，论文中除已注明部分外不包含他人已发表或撰写过的研究成果。对本文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。本声明的法律结果将完全由本人承担。

作者签名：  日期：2012 年 6 月 29 日

## 学位论文使用授权说明

本人完全了解哈尔滨工业大学关于保存、使用学位论文的规定，即：

（1）已获学位的研究生必须按学校规定提交学位论文；（2）学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（3）为教学和科研目的，学校可以将学位论文作为资料在图书馆及校园网上提供目录检索与阅览服务；（4）根据相关要求，向国家图书馆报送学位论文。

保密论文在解密后遵守此规定。

本人保证遵守上述规定。

作者签名：  日期：2012 年 6 月 29 日

导师签名：  日期：2012 年 6 月 29 日

# 致　谢

时间如白驹过隙，两年的硕士生活即将结束，值此论文完成之际，向所有关心、帮助过我的老师、同学、朋友表示衷心的感谢！

感谢哈工大社会计算与信息检索研究中心的所有老师，特别感谢中心主任也是我的导师刘挺老师。感谢刘老师给了我加入这个优秀集体的难得的机会，为我们提供了优越的工作环境和良好的学习科研氛围。感谢刘老师在学习工作中对我的指导和敦促，以及在为人处事的细节中教会我如何更好地成长。

感谢 UA 组张宇老师。在我两年的硕士生活中，张老师给予了我充分的信任和帮助。在张老师指导的 UA 组工作的这段时间，他渊博的学识、诲人不倦的教学态度以及平日里流露出的从容淡定、平易近人给我留下了深刻的印象。

从本科毕业实习半年到现在，在实验室两年多的时光是我人生路途中一段弥足珍贵的记忆。我和各位师兄师姐、同窗好友、师弟师妹们共同成长。

感谢宋巍师兄，从本科毕业设计到硕士论文的完成，都离不开他的帮助。他的严谨、勤奋和缜密的逻辑思维，都是我学习的榜样；感谢张伟男师兄，他的好学、热情、开朗乐观以及现在的减肥成功，都是我值得学习的。

感谢即将毕业的伍大勇师兄和已经毕业的赵静师姐、张文斌、康维鹏、韩中华等师兄，他们给予了我很多学习、科研上宝贵的经验。

感谢在实验室一起学习奋斗的郭江、唐国华、罗磊、高汉东、胡燊、王彪、唐都钰、陈炜鹏、张一博以及其他成员，谢谢你们平日里热心的帮助、信任和鼓励，愿大家前程似锦。

感谢师弟师妹们，愿你们学有所成，找工作顺利！

感谢两年多来“饭团”的陪伴，十几个人在学校餐厅共同就餐已经成为餐厅一道熟悉的风景。我会怀念与你们谈天说地时那份内心愉悦的心情。

感谢我的父母亲人，谢谢你们养育我长大。你们对我的爱，始终是我不断向前进取的动力。

感谢大家，愿好运与你们常伴！