基于用户行为及社交网络数据挖掘的智能信息推送系统

项目成员：武临风 侯一博 李欣 成晓东

杨华振 王建楠 莫奕海

指导老师：霍秋艳

所在院系：软件学院

所在班级：131014

项目时间：2013 年 5 月

目 录

摘 要

伴随着互联网的高速发展，我们每天所面临的信息量正在不断地增加，有时候我们甚至会因为信息量的过于巨大、繁杂而显得无从下手，最终选择匆忙浏览完成信息甚至放弃信息的获取。这便是信息爆炸时代人们所面临的共同的困难——信息越来越呈现出繁杂化和碎片化。

针对这个问题，项目组进行了充分的观察及调研，将问题总结为各网络站点用户体验不一、信息量冗余、无法智能提取信息三个点，并根据这三个点设计出了本项目——基于用户行为及社交网络数据挖掘的智能信息推送系统。该项目的最终目的在于为用户简单、快速、有效地获取到想要的信息。

Web 2.0 时代的一个显著特征就是社交网络逐渐充斥了人们的生活，人们越来越多热衷于在社交网络上进行个人及社会行为的表达及交流，这其中产生的社交数据包含了丰富的个人爱好、兴趣方向等信息，同时伴随着 Twitter、新浪微博/腾讯微博等社交媒体产品的出现，我们已经拥有到足够的社交网络信息量，让我们能够尽可能精准地对一个人的兴趣爱好进行分析、建模。同时，依赖于人们在互联网上积极的互动行为，我们能够让用户自发、主动地去进一步地对自身兴趣模型进行修正。最后，系统根据建立好的用户模型对互联网上的信息进行筛选、归类、排序，并按照一定地规则将互联网信息呈现给用户，让用户能够简单、快速、高效地获取到自己想要的信息，达到最佳的智能阅读体验。

关键词：社交网络数据挖掘　用户行为分析　阅读体验

一、 项目研究背景

1. 需求分析

随着互联网的高速发展，我们每天所面临的信息量也在不断增加，有时候我们甚至会因为信息量的过于巨大繁杂而显得无从下手，最终选择放弃信息的获取。这便是信息爆炸时代人们所面临的共同的困难——信息越来越呈现出繁杂化和碎片化。

针对如上问题，项目小组进行了充分调研，并总结结果如下：

- 各大网站网页结构不一，用户在切换使用的时候需要一定的时间对信息进行寻找、刷选、分析，使用户不能够快速获取到自己想要的信息；
- 同类型网站之间经常会产生大量重复的信息量（如网易新闻、搜狐新闻等），这将大量消耗用户的时间和精力；
- 网页中含有大量广告等无关信息，阻碍用户的正常阅读；
- 各大网站的功能及交互设计不一，给用户体验造成了一定的影响。

综上所述，在这个信息爆炸的时代，人们对于信息的“量”及“丰富程度”已经不再有过多的要求，但是在信息的“质量”及“是否对我有用”的需求上却是呼声愈高。如何快速、有效地获取到有价值的信息成为用户非常迫切的需求。

2. 研究背景

1) 社交网络数据挖掘

社交网络每天都会产生大量的用户数据（UGC，User Generated Content），并且具有空前的规模性和群体性，吸引着无数研究者从无序的数据中发掘有价值的信息。这就像概率统计中经常举的投硬币算其正反面概率的例子，从几次的投掷结果中很难看到规律，但通过几万次的大量投掷实验，便很容易看出正反面的出现次数几乎相等的规律。社交网络上产生了大量的规模化、群体化的数据，吸引了包括计算机科学、心理学、社会学、新闻传播学等领域专家和学者对其进行研究和探索，希望能够借助更强的社交网络的分析和处理能力发现更多人类尚未探

索出的规律。

对于社交网络的分析和研究范围很广，也存在着许多有意思的研究课题。例如，在社交网络中社区圈子的识别、社交网络中人物影响力的计算、信息在社交网络上的传播模型、虚假信息和机器人账号的识别、基于社交网络信息对股市、大选以及传染病的预测等。社交网络的分析和研究是一个交叉领域的学科，所以在研究过程中，通常会利用社会学、心理学甚至是医学上的基本结论和原理作为指导，通过人工智能领域中使用的机器学习、图论等算法对社交网络中的行为和未来的趋势进行模拟和预测。

2) 用户行为分析

用户信息行为指主体为满足某一特定的信息需求，在外部的刺激下表现出的获取、查询、交流、传播、吸收、加工和利用信息的行为。

就本质而言，用户信息行为具有以下一些主要特征：

- 信息行为是人类智力活动的产物，因而可以从认识论的角度加以研究。
- 信息行为为由信息心里活动决定，因而可以利用心理学理论方法研究信息心里-行为规律。
- 信息行为始终伴随着人的主体工作而发生，研究信息行为应与研究主体工作行为相结合。
- 信息行为是一种目的性很强的主动行为，对人的信息行为可以从总体上控制和优化。

在用户信息决策中，内驱力是由用户不断接受外界刺激后产生的一种信息内力，即现在的决策取决于用户过去接受刺激后的结果，如果行为导致好的结果，用户就有反复采取这种行为的趋势，否则就进行调节。用户的信息反应和行为除取决于刺激强度和诱因外，主要取决于习惯强度和内驱力。

二、 研究现状及研究意义

1. 研究现状

目前，国内外已有众多信息资讯订阅及智能推送站点，其设计理念及用户使

用体验各有特色，项目组选取了如下几个典型的信息资讯平台做了调研，调研结果如下：

- **无觅网**

无觅网是一个个性化阅读社区。在无觅网，用户可以挑选自己喜欢的话题，并得到关于这些话题的推荐内容；还可以关注感兴趣的其他用户，了解其他用户的阅读动态。无觅网利用人工智能技术根据用户阅读和喜欢过的内容，向用户推荐更多可能会感兴趣的内容（图 2-1）。

图 2-1

- **Google Reader**

Google 阅读器（Google Reader）是 Google 公司旗下一个基于网络的聚合器，能在线或离线阅读 Atom 和 RSS。英文版的 Google Reader 于 2005 年 10 月 7 日通过 Google 实验室发布，2007 年 9 月 17 日成为正式版。中文版的 Google 阅读器大约在 2007 年 9 月 18 日左右发布。可惜的是，作为 Google 第二个春季大扫除计划之一，Google 阅读器因用户数量逐年下降，将于 2013 年 7 月 1 日终止服务（图 2-2）。

图 2-2（来自维基百科）

- **新浪微博的“智能排序”**

新浪微博是由新浪网推出的，提供微型博客服务的类 Twitter 网站。用户可以通过手机客户端、网页等方式阅读或发布信息。随着微博信息的不断增多，为了能够给用户更好的阅读体验，新浪网为新浪微博推出了“智能排序”功能，其主要目的是为了帮助用户找到众多信息中最可能感兴趣的内容，提高阅读效率；同时通过排序及自动合并功能，帮助用户快速了解热点信息和好友共同的话题（图 2-3）。

图 2-3

2. 研究意义

从上述调研情况来看，尽管目前主流的资讯平台在提供基本的信息阅读功能之外都在努力为用户提供更好的阅读体验，但或因为其商业规划等原因未能将这些功能做到尽善尽美。

本项目组将致力于打造最简洁、方便、快速的智能信息推送系统，该系统将具有如下重要意义：

- **信息的获取方式将变得更加简单**

  用户只需通过新浪微博等社交网络账号进行登录，系统会自动获取其社交网络信息并将其进行用户兴趣建模，从海量的互联网信息中筛选出用户最感兴趣的部分信息并经过一定的整理排序后推送给用户。

- **个性化智能阅读信息推送**

  仅通过社交网络数据进行的建模往往会出现偏差，这时候用户就可以通过”喜欢”等方式对已建立模型进行一定的影响和改进，帮助系统进一步的了解用户，从而为用户更精准地推送信息。

- **与社交网络互动**

  用户可以通过”分享”的接口将信息分享到社交网络上与好友进行社交互动，促进用户阅读体验。

三、 研究目的及内容

1. 研究目的

如何为用户简单、快速、有效地获取到想要的信息是本项目的最重要的一个环节，同时这也是本项目组的最终研究目标。

但是在此之前，我们首先要能够知道用户需要的是什么样的信息，我们需要在获取信息之前对用户的个人兴趣爱好做一定的分析与建模，这时候我们就想到了是否能够利用用户的社交网络行为对其进行一定的用户兴趣分析？我们可以根据用户在新浪微博、腾讯微博等社交网络上的行为及关注点信息，对用户兴趣爱好及关注点进行建模，根据所建立模型对外界信息进行筛选、整理及排序，并

推送给用户，用户仅需打开应用即可获取到 TA 所想要获取到的信息。

但是基于社交网络信息所分析出的用户行为总会存在一定的偏差，于是我们同样基于用户行为分析技术设计了一套新的功能体系——“喜欢/分享”。该功能体系重在让用户能够自主权衡信息权重，从而让我们之前的建模能够做到更充分精确。

基于上述理念，最终我们将实现一套基于社交网络数据挖掘、用户行为分析技术的信息推送系统。

2. 研究内容

我们的最终目的是实现一套基于社交网络数据挖掘、用户行为分析技术的信息推送系统，并将本项目研究内容的业务流程图拟绘如下（图 4-1）：

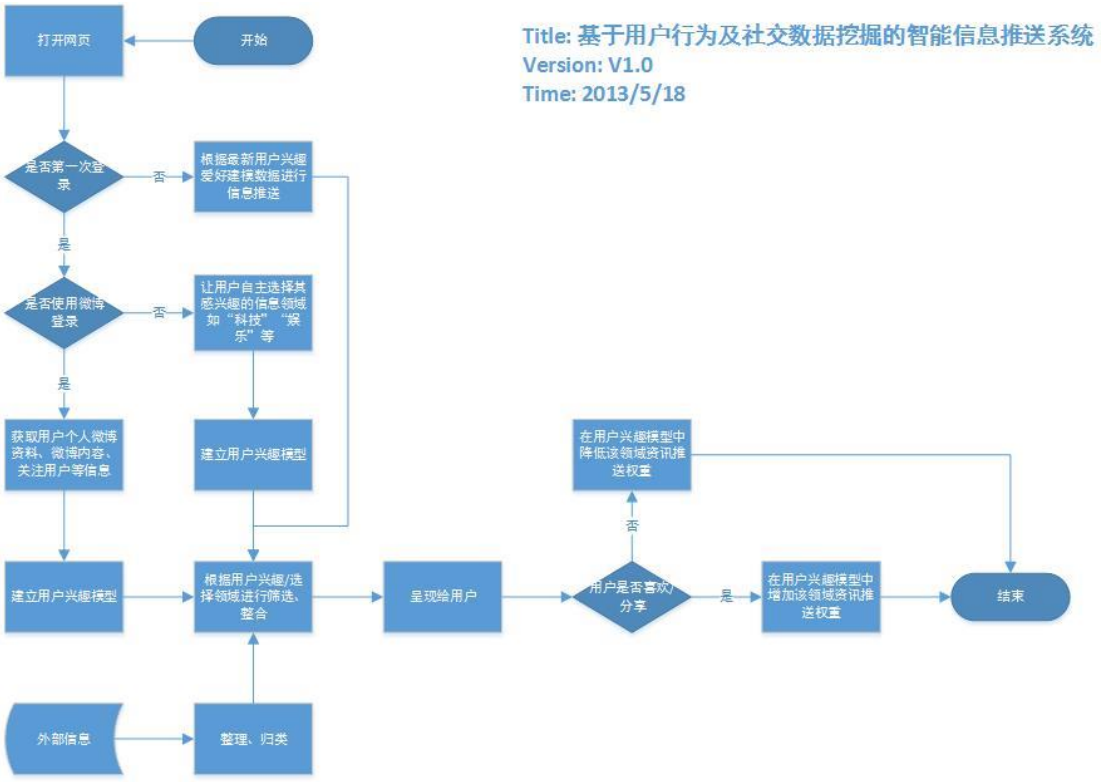

图 4-1

项目整体架构图如下（图 4-2）：

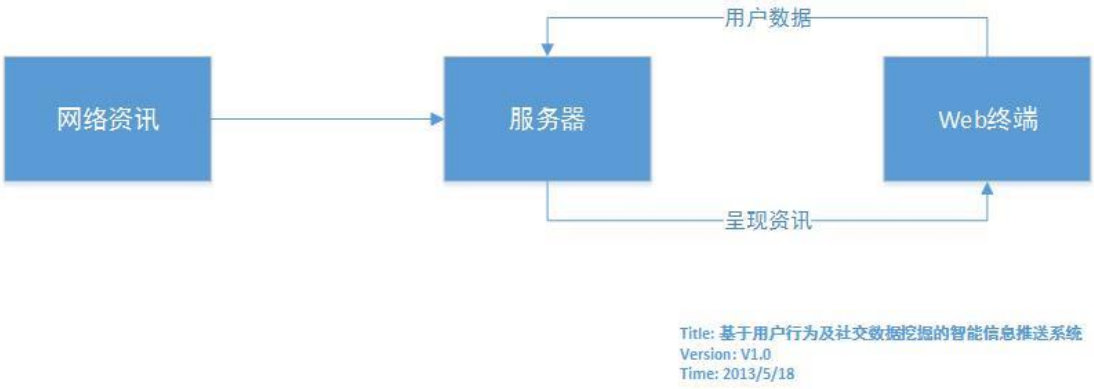

图 4-2

具体研究内容如下：

- **社交网络数据挖掘**

在用户使用微博账号进行登陆后，获取用户的个人资料，标签，发布的微博内容以及关注的用户等信息，对用户的社交网络进行数据建模，并通过语法以及语义分析对提取到的微博内容进行统计语言模型建模，使用数据挖掘算法从数据模型中提取出用户可能感兴趣的关键词并赋予权重，若词汇表(数目为 N)的关键词在该用户的提取出的关键字中没有出现则对应的权重为 0。词汇表所有的关键字构成了一个 N 维向量来对该用户进行描述。

- **用户行为分析**

用户在阅读文章后会标记是否喜欢或者选择是否分享，通过用户这一信息行为对该用户相关的关键字的权重进行调整。如果用户分享文章或者标记文章，则说明该用户对该关键字的信息表现出了利用，吸收，传播的需求，即提高该关键字的权重。反之说明用户对该关键字的信息反应并不强烈，需要适当降低该关键字的权重。

- **推送资讯的筛选及排序**

首先，使用向量对新闻进行描述。对一篇新闻中的所有实词，计算出它们的文本词汇频率/逆文本频率值（TF/IDF）。我们按照这些实词在词汇表的位置对它们的 TF/IDF 值排序，如果单词表中的某个词在新闻中没有出现，对应的值为零。词汇表中所有词(数目为 N)的 TF/IDF 值组成一个 N 维的向量。我们使用该向量来刻画这篇新闻。

然后，使用余弦定理计算两篇新闻的 N 维向量。当两条新闻向量夹角的

余弦等于一时，这两条新闻完全重复，使用此方法来删除重复的网页。

最后，使用余弦定理计算新闻向量与用户向量的余弦值，当余弦值大于阈值后，将该新闻推荐给该用户。余弦值越高，排名越靠前。

- **前端设计**

前端采用 BootStrap 框架进行设计，以快速开发出以 Flat UI 为主体设计风格的站点，支持所有现代浏览器，并支持跨平台使用。

四、 预期成果

综上所述，项目组在本项目完成后将达到以下预期成果：

1. 基于用户行为及社交网络数据挖掘的智能信息推送系统（Web 端）；
2. 完成用户行为及社交网络数据挖掘的研究报告；
3. 在公开学术期刊发表相关论文。