同济大学
TONGJI UNIVERSITY

硕士学位论文

(专业学位)

一种基于 P2P 网络的信息自主流动机制的研究

姓　　名：　先毅昆

学　　号：　1236191

所在院系：　软件学院

职业类型：

专业领域：　软件工程

指导教师：　张晨曦

副指导教师：奚自立

二〇一五年三月

同濟大學
TONGJI UNIVERSITY

A dissertation submitted to

Tongji University in comformity with the requirements for

the degree of Master of Computer Science

Study of the Mechanism of Information Spontaneous Propagation on P2P Network

Candidate:	Yikun Xian
Student Number:	1236191
School/Department:	School of Software Engineering
Discipline:	
Major:	Software Enginnering
Supervisor:	Chenxi Zhang

March, 2015

一种基于P2P网络的信息自主流动机制的研究

先毅昆 同济大学

学位论文版权使用授权书

本人完全了解同济大学关于收集、保存、使用学位论文的规定，同意如下各项内容：按照学校要求提交学位论文的印刷本和电子版本；学校有权保存学位论文的印刷本和电子版，并采用影印、缩印、扫描、数字化或其它手段保存论文；学校有权提供目录检索以及提供本学位论文全文或者部分的阅览服务；学校有权按有关规定向国家有关部门或者机构送交论文的复印件和电子版；在不以赢利为目的的前提下,学校可以适当复制论文的部分或全部内容用于学术活动。

学位论文作者签名：

年　　月　　日

同济大学学位论文原创性声明

本人郑重声明: 所呈交的学位论文, 是本人在导师指导下, 进行研究工作所取得的成果。除文中已经注明引用的内容外, 本学位论文的研究成果不包含任何他人创作的、已公开发表或者没有公开发表的作品的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体, 均已在文中以明确方式标明。本学位论文原创性声明的法律责任由本人承担。

学位论文作者签名：

年　　月　　日

摘　　要

应概括地反映出本论文的主要内容, 包括工作目的、研究方法、研究成果和结论, 要突出本论文的创造性成果。摘要力求语言精炼准确, 硕士学位论文建议 1000 字以内, 博士学位论文建议 3000 字以内。摘要中不要出现图片、图表、表格或其他插图材料。

学位论文原则上应用汉语撰写, 对于用汉语授课并享受中国政府奖学金的博士硕士留学研究生, 学位论文如用英语 (德语、法语) 撰写, 硕士学位论文不少于 3000 汉字摘要, 博士学位论文不少于 5000 汉字摘要; 对于其他情况 (含用英语授课) 的博士硕士留学研究生, 学位论文如用英语 (德语、法语) 撰写, 可不要求撰写汉语摘要, 但必须有英语摘要。

关键词是为了便于文献索引和检索工作, 从论文中选取出来用以表示全文主题内容信息的单词或术语, 摘要内容后另起一行标明, 一般 3 5 个, 之间用“,”分开。

关键词：关键字，摘要

Abstract

First Paragraph

SecondParagraph

Tongji University, located in Shanghai, has more than 50,000 students and 8,000 staff members (as of 1 September 2007). It offers degree programs at both undergraduate and postgraduate levels. Established in 1907 by the German government together with German physicians in Shanghai, Tongji is one of the oldest and most prestigious universities in China. Among its various departments it is especially highly ranked in engineering, among which its architecture, urban planning, and civil engineering departments have consistently ranked first in China for decades, and its automotive engineering, oceanography, environmental science, software engineering, German language departments are also ones of the best domestically.

Kew Words: English Keywords, Abstract

Contents

第 1 章 绪论

1.1 选题背景

互联网技术的蓬勃发展在很大程度上给人们的生活带来了越来越多的便利，人们在逐渐适应网络这样的平台的同时，也更倾向于甚至依赖在网络平台上完成生活中的各种事情，可以说，网络对于人们的重要性已经几乎等同于空气、水和食物。按照网络平台的功能来划分，门户网站（新浪、搜狐等）是新闻实事评论发布的主要渠道，社交网站（微博、人人等）是人们分享个人想法和心情的首选，博客系统（新浪博客、百度空间等）是人们发表和传达思想和经验的主要平台，以及一些新新涌现的图片和视频等多媒体资源分享平台（优酷土豆、POCO等）大大丰富了人们的娱乐生活。此外，鉴于国内发达的物流行业，各大电子商务平台（淘宝、京东等）也使得不用出门就能购物成为了现实。按照网络平台资源的载体来划分，包括文本、图片、音乐和视频几大类，其中文本无疑是整个互联网资源的主体，无论是传统的新闻、博客、评论、说明，还是新新发展的弹幕视频网站（Acfun[1]、Bilibili[2]等），都是由很多文本信息组成的。按照网络平台的用户参与角度来划分，可以将用户角色分为两种：信息发布者和信息获取者。以程序员为例，当他需要学习一项新技术时，往往会通过搜索引擎寻找一些与该技术相关的教学经验文章，从而使自己尽快掌握该项技术。等对该技术数量掌握之后，往往又会通过写技术博客的方式记录下他的学习历程和使用中的经验之谈，以供他人参考。同时，该用户还肯定拥有其他多个兴趣爱好，比如他可以与网上其他用户分享旅游游记、摄影作品等等。

虽然这些网站系统在功能上、信息载体上或是用户参与方式上都截然不同，但是宏观来看他们都存在一个共同的问题：信息孤岛现象。如图 1.1所示，当这些网站发展越来越多之时，各个网站之间信息不流通的问题也日渐明显，每个网站独立发展，出于安全性等方面的考虑，其资源和用户的数据并不能实现跨平台共享，从而使得每个网站成为一座座信息孤岛。

举例来说，当用户作为信息获取者时，虽然每个网站可以为本平台上的用户提供很好的用户体验，通过数据挖掘等技术发现用户的兴趣，为其推荐有潜在需求的信息，但是不同网站之间的用户兴趣不能共享，导致兴趣推荐的不准确。当用户作为信息发布者时，其操作会更加繁琐，他往往需要打开多个网站重复几乎相同的操作来发布同样的内容，最明显的证据就是同时使用微信、微博和人人的用户需要在三个平台上重复三次操作完成发布。当这些网站越来越多的时候，问题也随之而来，即用户可能需要打开很多个独立的网站来完成一系列类似的事情。比如，先在新闻网站上浏览最新发生的时事，接着在摄影网站上发布新照

[1]http://www.acfun.tv/
[2]http://www.bilibili.com/

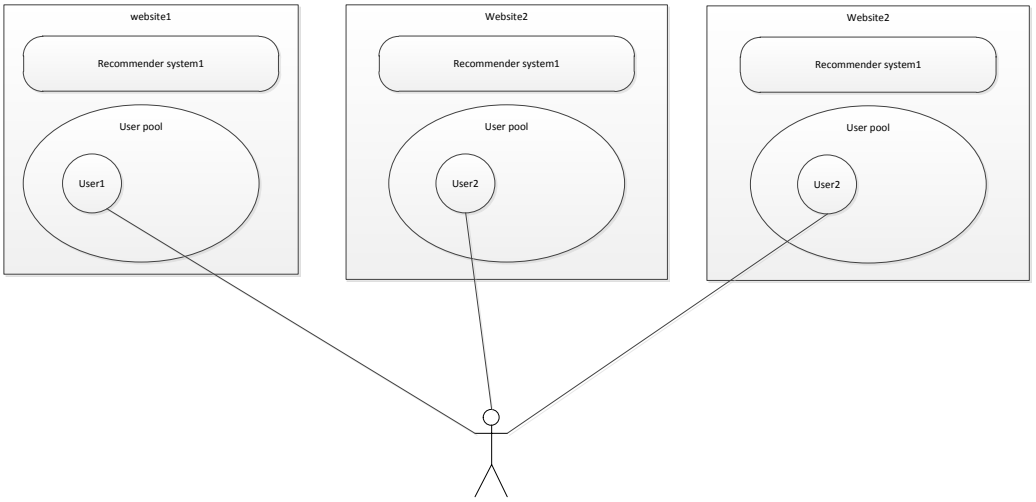

图 1.1: 互联网平台的信息孤岛现象

片，然后在社交网站上浏览好友更新的状态，最后在电商网站上购买一些商品等等。换句话说，用户是单向地去寻找想要的信息，这其中无疑包含了一些不必要的重复操作。退一步来说，即使存在某些用户只在社交网站上浏览信息，很少接触其它网站，也会有一些重复操作的问题。因为该用户的好友很有可能是分散地活跃于各个不同的社交网站平台 (微博、人人网等)，而且每个平台发布的信息肯定有所不同，所以该用户仍然需要逐个登录各个网站后才能浏览到各个用户的信息。再退一步说，即便现在已经有些软件把所有社交网站整合成一个统一接口，让用户只需一次登录就能同时访问多个平台，用户仍然会遇到信息冗余的问题，比如重复的新闻、不感兴趣的推荐等等。

为了解决上述问题，我们设想有这样一个智能系统：每当用户打开系统时，系统会自动推送今天的时事新闻、其好友最近更新的状态、感兴趣或者正在促销的商品，以及一些根据用户偏好过滤的信息。此外，他也可以在系统上发布自己的信息给他的好友，甚至给那些对他信息感兴趣的陌生人。虽然，实现这样一个系统的工作量和难度是巨大的，但仔细观察后可以发现这样一个重要的规律：即用户希望所有的信息能够在整个互联网上智能地自主流动，在用户单方面寻找信息的同时，让信息也能自主地流向符合特定需求的用户。

从抽象层面来看，要实现这样一套信息自主流动的机制，传统的集中式计算模式已经不再适用。如图 1.2所示，这里每个用户均看作一个独立的节点，所有的节点整合在一起就形成一个巨大的 P2P 网络。其中，每个节点既充当服务器用于分发信息，也充当客户端用于接收信息，并且由某个节点发出的信息在其它节点间传播的时候会自主流动，寻找潜在的、匹配的节点。简单来说，要使信息

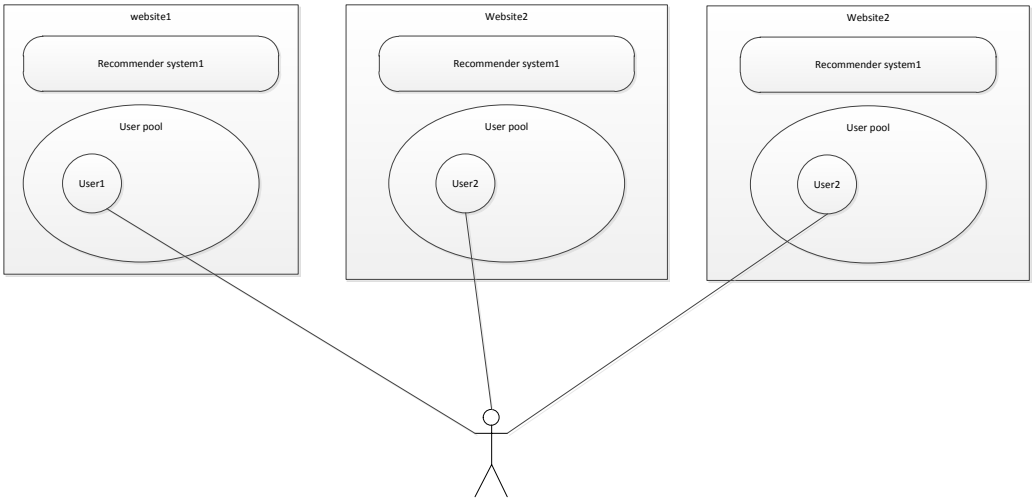

图 1.2: 一种基于 P2P 网络的智能系统解决方案

能够自主流动，就要攻克这几方面的难点：

- 资源信息描述与匹配：互联网资源普遍呈现出半结构化或非结构化的特征，以普通的新闻或者博客为例，除了作者、标题、时间等结构化信息，正文纯文本都是典型的非结构数据。因此，如何从这些非结构化的文本信息中提取出有用的特征、并用这些特征来衡量这些文本之间的相关性是本文重要研究内容。

- 用户兴趣挖掘与关联：用户的兴趣可以从其发布和查看的信息中反映出来，基于上述对资源信息的描述模型，可以训练出某一时刻或者某一时段内的用户兴趣。但是，由于用户兴趣可能会随着时间的推移而不断变化，并且在同一时刻可能同时存在长期兴趣和短期兴趣。因此，这部分需要基于资源模型来解决用户兴趣随时间变化的问题。

- P2P 网络路由与搜索：从整个网络结构来看，在初始状态下各个用户节点之间的关系仅仅为物理意义上的距离关系，换句话说，节点与节点之间的边的权重不能代表用户与用户之间兴趣的相似度，从而也就导致每个节点上的信息不能自由传播。因此，这部分需要借助用户兴趣模型，通过 P2P 网络的路由搜索算法来寻找潜在的相似节点，最后构成某一时间段内的稳定结构。

1.2 研究意义与应用价值

对于上述提出的信息孤岛现象是主要由目前网站之间相互独立而导致的，这些中心化结构的独立网站架构存在以下几个问题：

- 从资源传播的角度来看，每个网站的数据都仅仅在一个局部范围内可用，但是用户却要在多个这样的局部范围内同时使用多个网站。虽然对于每个网站可以维护相对较好的资源整合和管理，但是对于属于用户的资源的传播带来了很大的阻碍。
- 从计算资源的角度来看，每个独立的传统 B/S 架构的网站通常由背后的一套计算群组来支撑，当网络流量十分巨大的时候，服务器的硬件和软件性能决定了系统平台可承受的最大程度。而在用户这边的客户端的 I/O 消耗相对小很多，但是大量客户端的请求对于服务器是一个不小的考验，从而导致了服务器成为整个系统的计算瓶颈。
- 从隐私安全的角度来看，目前网站的数据均由平台自己的数据库进行维护，对于数据泄露成为重大隐患之一，即使网站的安全性是非出色，用户也会担心其数据是否会被第三方利用于商业用途。
- 从拓扑架构的角度来看，用户与网站系统组成了中心化的拓扑结构，该结构不利之处在于当网站服务器宕机或者网络中断的时候，所有用户均无法访问该系统，这就使得网站系统的备份机制与主从切换机制十分完善，而对于大部分初创公司的流量有限和资金不足等特征是一个很大的考验。

与中心化结构网站架构不同，在基于 P2P 网络架构中，每个节点都可以既可以充当服务器为其它节点提供资源，也可以作为客户端向其它节点获取资源。综合来说，P2P 网络的应用具有以下几大优点 [1]：

- 非中心化（Decentralization）：在 P2P 网络中，节点之间的通信无需经过中心服务器，因此大大降低了服务器因为计算资源问题而导致的性能瓶颈。同时，网络中的信息资源完全存储在各个节点中，信息的交互完全在节点之间进行。
- 可扩展性（Scalability）：由于 P2P 网络中的每个节点可以看做服务提供者，所以每当有新的节点加入时，相当于扩充了整个网络的服务能力和资源，相反用户节点也可以随时退出，十分的灵活。以网络下载资源为例，普通 C/S 模式的计算下载通道的传输能力基本依赖于服务器的带宽，当用户一旦很多的时候，下载速度往往会受到很大的影响，但是在 P2P 网络中却恰恰相反，更多的节点请求意味着有更多的资源提供者，从而下载速度也更快。
- 健壮性（Robustness）：与中心化架构模式不同，P2P 网络不会因为某一个节点退出而在很大程度上影响网络的正常运行，相反只有很小的一部分影响。结构良好的 P2P 网络一旦在某几个节点发生异常的时候会自动调整拓扑结构，从而保证整个网络的连通性。

- 安全性（Security）：传统集中式的计算模式在通信上往往需要经过中心服务器，也就是说用户的数据信息会全部经过这些服务器，因此只要在这些为数不多的服务器上动些手脚，用户隐私就会被泄露。但是，在 P2P 网络上则不同，信息传输不会经过某些特定的节点，并且还可以与特定有需求的目标节点建立连接，从而在很大程度上提升了匿名通信的可靠性以及灵活性。
- 负载均衡（Load-Balance）：根据上文所述，P2P 网络下的用户节点既能充当信息发布者的角色，也可以作为信息接受者的角色，所以在计算资源和存储资源的分布十分均匀。同时，对于信息接受者来说，其可以向计算资源空闲的节点请求资源，而不用排队等待那些繁忙的节点直到任务完成。

基于上述有点，本文首先建立一套用于描述互联网信息和用户行为偏好的模型，旨在将这些具体的信息和用户行为偏好上升到抽象层面，剥离出一个统一的输入输出接口。对于那些专注于复杂的数据分析和建模工作的模块，该接口起到了整合的作用。而且，在此之后，也可以展开专门基于该描述模型的信息与用户匹配算法的研究。其次，根据上述信息描述模型与用户行为偏好模型的特点，从拓扑结构，通信机制等方面入手，建立一套合适的分布式架构 (如 P2P 的计算模式)，并利用上述的匹配算法，研究由各个节点组成的覆盖网络间的信息自主流动的机制。最后，将上述两项内容结合在一起，实现一套完整的信息自主流动的原型系统。

1.3 国内外研究现状

从目前的科学研究和商业应用方面来看，还没有关于一套基于 P2P 网络的信息自主流动机制的相关工作。因此，这部分将从文本信息挖掘、用户兴趣关联以及 P2P 网络搜索等三方面来具体展开阐明已有的研究工作。

1.3.1 文本挖掘相关工作

在文本信息挖掘方面，由于当今的硬件和软件技术在各方面都有大幅度提高，产生了大量的不同类型的数据资源 [2]，特别是纯文本的数据。这些文本大多来自于社交网络、新闻博客等网站，内容形式对于人们来说也是十分的丰富。但是这对机器来说却没有那么容易识别，因此需要设计一系列的算法来从这些文本数据中找出一些模式和规律，从而让机器更好地利用这些数据为人们创造更大的价值。结构化数据（Structured）通常可以存储在各类数据库中，但是纯文本数据却因为它的半结构化（Semi-structured）或者非结构化（Unstructured）特性从而只能有通过搜索引擎等技术才能进行索引和查找 [3]。搜索引擎是信息检索（IR）中的一种方式，其作用是方便用户直接通过输入关键词找到内容相关的文

档集合，其目标是如何通过有效和高效的方法是检索的结果更加精确，涉及的研究领域包括文本聚类、文本分类、文本概括、推荐系统等 [4, 5, 6]。但是，在本文中，除了这些信息检索基本的需求，更重要的是从文本中挖掘出重要的特征和模式，将原本非结构化的数据量化成半结构化数据，并且这种量化过程更多地从用户兴趣这点切入的。针对这个要点，文本挖掘技术一般可以分为以下几大类：

1.3.1.1 文本抽取

文本抽取（Text Extraction）主要是从半结构化或结构化的纯文本里找出结构化信息，涉及了自然语言处理（NLP），信息检索和 Web 挖掘（Web Mining）等领域的技术。该研究最基本的两大工作是：a）命名实体检测（NER），包括找出文本中的人物、地点、组织等等；和 b）关系抽取（Relation Extraction），主要包括名字之间的地理位置关系、人物关系、动作执行关系等等。

其中，对于命名实体检测问题，主要包括基于规则的方法 [7, 8] 和基于统计学习的方法，后者有三个十分著名和常用的方法：隐马尔可夫模型（HMM）[9]，这是一个简单且有效的生成模型（Generative Model），最大熵马尔可夫模型（MEMM）[10]，这是一个判别模型（Discrimitive Model），适合于训练集十分充足的情况，因为它能给出一个更小的预测误差，以及条件随机场（CRF）[11]，该模型是一个基于无向图的判别模型，因此适用于当前状态均受前后影响的情况。

对于关系抽取问题，主要包括基于分类的方法（Classification-based）[12, 13, 14, 15] 和基于核的方法（Kernal-based）[16, 17]。其中，分类方法在应用之前需要对文本进行特征提取（Feature Extraction），比较有用的特征包括实体、词汇、语法和背景知识等。而核方法最常见的场景是先将文本用向量空间模型（VSM），再进行下一步计算，主要分为基于序列的核方法、基于树的和方法和混合核方法，在机器学习中，核函数（Kernal Function）是两组向量空间的内积，也可以看做观测值之间的一种相似度衡量方法，从而观测值不需要显示地映射成向量空间。

总结来说，当前命名实体检测最有效的方法都是基于概率图模型而提出的，典型代表是 MEMM 和 CRF，而关系抽取的方法则是很大程度上基于选取的特征或者定义的核函数，然后再利用分类算法进行解决。可以发现，上述这些主流方法都是有监督方法，但是，随着互联网产生数据的规模不断增大，半监督和无监督方法会更加实用。

1.3.1.2 文本摘要

文本摘要（Text Summarization）用于从纯文本中识别出重要的内容，包括重要的句子、关键词和主题等。因此根据不同的表示方式，基本方法可以分为主题

表示法、标识符表示法、句子表示法。

其中，主题表示法需要首先找出一种主题描述的方式，一般是用主题词，这种词的特性是排除文档中出现频率高以及极少出现的词，然后利用对数似然比检验（LR Test）来识别那些能够很具有代表性的单词 [18]。关于单词出现频率的计算方式，最经典的就是 TF-IDF[19]，这种方法一方面计算了单词出现的频率，另一方面也衡量了单词的重要性。再进一步，文献 [20] 提出的潜语义分析（LSA）方法则用一种隐含的方式来表示文本中的寓意信息，其十分适用于对新闻文本的总结，而且不需要使用第三方的词典数据就能实现。在 LSA 的基础上，更多的基于贝叶斯理论的主题模型被陆续提出，这些模型在主题表示上十分成熟，而且很多已经在工业界广泛使用 [21, 22, 23, 24]。

对于句子粒度的摘要方法，通常是先将相似的句子进行聚类 [25, 26]，其中相似度的计算方式包括简单的余弦相似度（Cosine Similarity）等。处于同一类中的句子被认为是某种主题，而有很多句子的类则说明包含了重要的主题，然后从每个这样的类中选取一个句子作为这种主题的代表，组合起来就是对文章的摘要。同时，对于选出来的句子可以用 KL 散度（KL Divergence）来进行评分，从而来判断选出来的摘要的好坏。

1.3.1.3 文本降维

文本降维（Dimensionality Reduction）用于解决文本向量的维度过大的问题，这里主要的方法包括诸如 LSI（Latent Semantic Indexing）、PLSI（Probabilistic Latent Semantic Indexing）和 LDA（Latent Dirichlet Allocation）[27, 28] 等主题模型，这些模型共同的假设前提是把文档看做词袋模型（Bag-of-words），即不关心单词的先后顺序，从而可以更有效更快速地来处理文本。

对于主题模型来说，在整个文档集合 D 中，共有 M 篇文档，词汇表共有 W 个不同的词汇，并且假设总共有 K 个主题，其中文档和词汇都是已知的，主题为待求的问题。因为已经存在 $D*W$ 维的文档 -词汇矩阵 X，通过主题模型以后本质是将该矩阵分解成两个矩阵：$D*K$ 维的文档 -主题矩阵 A，和 $K*W$ 维的主题 -词汇矩阵 B，即 $X = A*B$。在矩阵 X 中，每一行表示一篇文章 d_m，每一列表示一个词汇 v_i，每个值 $x_{m,i}$ 表示该文档具有该词汇的比重，这个比重可以是通过正规化后的词频，也可以是 TF-IDF 的值。在矩阵 A 中，每一行表示文档 d_m，每一列表示主题 z_k，每个值 $a_{m,k}$ 表示该文档具有该主题特征的权重，所以文档可以表示为一个关于主题的 K 维向量，从而将文档看成一个服从多项分布的随机变量 $\vec{\theta_m}$。在矩阵 C 中，每一行表示主题 z_k，每一列表示词汇 v_i，每个值 $b_{k,i}$ 表示该主题中该词汇占有的比重，由此可见，这里的每个主题表示为关于词汇的一个 W 维向量，从而主题可以看做一个服从多项分布的随机变量 $\vec{\phi_k}$。关

于如何求出参数 $\vec{\theta_m}$ 和 $\vec{\phi_k}$ 将在后文结合本文提出的模型具体详解。

但是诸如 LDA 这类的主题模型还有很多局限性，第一，LDA 推理的方法大致分为两种 Collapsed Gibbs sampling[28] 和 Variational Approximation[27]，但是两者在整个迭代过程中十分耗时，当数据量很大的时候，特别是词汇表的大小 W 达到百万级的时候，$\vec{\phi_k}$ 的维度十分高，导致计算很慢。因此，基于这个问题，目前已经提出了并行的亦或是分布式的 LDA 算法 [29, 30]，从而应对工业界大规模数据的问题。第二，当文档集随时间变化时，LDA 本身并不能很好适应这个动态的过程，一种方法是引入时间变量重构概率图模型。文献 [20] 通过扩展了时间维度和 LSI 的张量分析从而实现了对历史文档的更新。文献 [31, 32] 则在 PLSI 的基础上加上时间维度提出两个全新的主题模型，分别在视频中活动抽取和音乐抽取方面有所作为。David Blei 作为 LDA 的提出者，后来根据时间动态变化这个需求又提出了基于时间演化的主题模型 [33]，很好地解决了这方面的问题。第三，有些文档间会存在关联，比如学术论文间可以以作者和引用上建立连接，在诸多改进方案之中，RTM 模型（Relational Topic Model）[34] 在 LDA 的基础上联合模拟文档和链接的生成过程，从而很好地解决了文档网络的问题。

总结来说，基于概率图的主题模型是一种很好的文本降维方式，一方面它以统计学中的降维方法为基础，使得我们可以推理或近似得出隐藏在文档背后的主题是什么，另一方面，它也给出了对主题的一种新的定义，并且基于这种定义使得模型能够很容易地根据不同类型文档的需求来改进。但是一个很重要的问题是如何能够将这些主题模型应用到实际工业界中，因为在性能、效率方面还是有不尽如人意的地方。

1.3.2 用户兴趣挖掘相关工作

1.3.2.1 用户情感分析

1.3.3 P2P 网络路由搜索相关工作

1.4 本文内容安排

本文的主要内容安排如下。首先本章对该研究的背景，主要研究内容和研究意义与研究现状进行了一个总体的介绍。第二章将对基于 P2P 网络的信息自主流动机制提出一个完整的技术方案，并且结合研究现状和存在的问题对设计需求和挑战进行详细描述，从而给出关键的技术点。第三章将对解决方案中的第一部分即文本资源的描述与匹配问题进行深入分析，通过结合现有成熟的模型，提出一整套全面的理论模型。第四章将结合第三章的研究基础对用户兴趣进行建模，结合现实中兴趣的动态变化特征给出相应的解决方案。第五章将基于第四章的研究基础提出一种基于 P2P 网络的路由与搜索算法，从而实现动态构建兴

趣覆盖网络。第六章将根据第三、四、五章的理论进行模拟实验和真实实验，验证上面提出的新的方法的有效性。第七章则基于前文的研究成果实现了一套简单的智能系统原型，包括功能设计和系统架构等方面。最后第八章对本研究做了一个总结, 并对研究进一步的工作方向进行了讨论。

第 2 章　信息自主流动机制的方案概述

本章将阐述基于 P2P 网络的信息自主流动机制的基本方案。从用户和互联网传播信息的关系入手，根据信息本身在抽象层面上的通用属性以及用户的行为和偏好，建立一套统一的信息描述模型和用户行为偏好模型。其次，基于已建立的两种描述模型，利用相关的算法，进行信息间匹配问题的研究。然后利用信息传播中起点、终点和传播路径等参数的特点，建立一套分布式系统的拓扑结构，并利用信息描述模型，整合成一套完整的系统逻辑结构模型。最后，针对该结构模型的特征和属性，研究各个传播节点间的通信模式，最终形成一套完整的信息自主流动机制。

2.1　文本信息的匹配与推荐模型

纯文本数据的特征：稀疏性和高维度，举例来说：given corpus may be drawn from a lexicon of about 100,000 words, but a given text document may contain only a few hundred words. Thus, a corpus of text documents can be represented as a sparse term- document matrix of size n × d, when n is the number of documents, and d is the size of the lexicon vocabulary. The (i,j)th entry of this matrix is the (normalized) frequency of the jth word in the lexicon in document i. The large size and the sparsity of the matrix has immediate implica- tions for a number of data analytical techniques such as dimensionality reduction.

2.2　用户兴趣的挖掘与描述模型

2.3　P2P 网络的路由与发现技术

2.4　信息自主流动的原型系统

2.5　小结

第 3 章　互联网文本信息的匹配与推荐模型研究

3.1　小结

第 4 章　用户兴趣模式的挖掘与描述模型的研究

4.1　小结

第 5 章　基于 P2P 网络的信息自主流动机制的基础架构

5.1　小结

第 6 章　实验分析与结果分析

6.1　小结

第 7 章　原型系统设计与实现

7.1　小结

第 8 章　结论与展望

致谢

逾尺的札记和研究纪录凝聚成这么薄薄的一本, 高兴和欣慰之余, 不禁感慨系之。记得鲁迅在一篇文章里写道:“人类的奋战前行的历史, 正如煤的形成, 当时用大量的木材, 结果却只是一小块”。倘若这一小块有点意义的话, 则是我读书生活的最好纪念, 也令我对于即将迈入的新生活更加充满信心。回想读书生活, 已经整整二十个年头, 到同济求学将近五年, 攻读博士学位也已三年了。进入同济大学以来, 深深醉心于一流学府的大家风范。名师巨擘, 各具特点; 中西融合, 文质相顾。处如此佳境以陶铸自我, 实乃人生幸事。

2015 年 3 月

参考文献

[1] 罗杰文, ``Peer to peer (p2p) 综述," 2005.

[2] J. Han, M. Kamber, and J. Pei, *Data mining, southeast asia edition: Concepts and techniques*. Morgan kaufmann, 2006.

[3] W. B. Croft, D. Metzler, and T. Strohman, *Search engines: Information retrieval in practice*. Addison-Wesley Reading, 2010.

[4] D. A. Grossman, *Information retrieval: Algorithms and heuristics*, vol. 15. Springer Science & Business Media, 2004.

[5] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*, vol. 1. Cambridge university press Cambridge, 2008.

[6] G. Salton and M. J. McGill, ``Introduction to modern information retrieval," 1983.

[7] D. E. Appelt, J. R. Hobbs, J. Bear, D. Israel, and M. Tyson, ``Fastus: A finite-state processor for information extraction from real-world text," in *IJCAI*, vol. 93, pp. 1172--1178, 1993.

[8] R. Mooney, ``Relational learning of pattern-match rules for information extraction," in *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, pp. 328--334, 1999.

[9] R. Dugad and U. B. Desai, ``A tutorial on hidden markov models," *Signal Processing and Artificial Neural Networks Laboratory Department of Electrical Engineering Indian Institute of Technology*, 1996.

[10] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, ``A maximum entropy approach to natural language processing," *Computational linguistics*, vol. 22, no. 1, pp. 39--71, 1996.

[11] J. Lafferty, A. McCallum, and F. C. Pereira, ``Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[12] N. Kambhatla, ``Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations," in *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, p. 22, Association for Computational Linguistics, 2004.

[13] Z. GuoDong, S. Jian, Z. Jie, and Z. Min, ``Exploring various knowledge in relation extraction,'' in *Proceedings of the 43rd annual meeting on association for computational linguistics*, pp. 427--434, Association for Computational Linguistics, 2005.

[14] J. Jiang and C. Zhai, ``A systematic exploration of the feature space for relation extraction.,'' in *HLT-NAACL*, pp. 113--120, 2007.

[15] Y. S. Chan and D. Roth, ``Exploiting background knowledge for relation extraction,'' in *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 152--160, Association for Computational Linguistics, 2010.

[16] R. C. Bunescu and R. J. Mooney, ``A shortest path dependency kernel for relation extraction,'' in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 724--731, Association for Computational Linguistics, 2005.

[17] S. Zhao and R. Grishman, ``Extracting relations with integrated information using kernel methods,'' in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 419--426, Association for Computational Linguistics, 2005.

[18] T. Dunning, ``Accurate methods for the statistics of surprise and coincidence,'' *Computational linguistics*, vol. 19, no. 1, pp. 61--74, 1993.

[19] S. Gupta, A. Nenkova, and D. Jurafsky, ``Measuring importance and query relevance in topic-focused multi-document summarization,'' in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pp. 193--196, Association for Computational Linguistics, 2007.

[20] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, ``Indexing by latent semantic analysis,'' *JAsIs*, vol. 41, no. 6, pp. 391--407, 1990.

[21] H. Daumé III and D. Marcu, ``Bayesian query-focused summarization,'' in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 305--312, Association for Computational Linguistics, 2006.

[22] A. Haghighi and L. Vanderwende, ``Exploring content models for multi-document summarization,'' in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 362--370, Association for Computational Linguistics, 2009.

[23] D. Wang, S. Zhu, T. Li, and Y. Gong, ``Multi-document summarization using sentence-based topic models," in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pp. 297--300, Association for Computational Linguistics, 2009.

[24] A. Celikyilmaz and D. Hakkani-Tur, ``A hybrid hierarchical model for multi-document summarization," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 815--824, Association for Computational Linguistics, 2010.

[25] K. McKeown, J. Klavans, V. Hatzivassiloglou, R. Barzilay, and E. Eskin, ``Towards multidocument summarization by reformulation: Progress and prospects," in *AAAI/IAAI*, pp. 453--460, 1999.

[26] V. Hatzivassiloglou, J. L. Klavans, M. L. Holcombe, R. Barzilay, M.-Y. Kan, and K. McKeown, ``Simfinder: A flexible clustering tool for summarization," 2001.

[27] D. M. Blei, A. Y. Ng, and M. I. Jordan, ``Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993--1022, 2003.

[28] G. Heinrich, ``Parameter estimation for text analysis," tech. rep., Technical report, 2005.

[29] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh, ``On smoothing and inference for topic models," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 27--34, AUAI Press, 2009.

[30] A. Smola and S. Narayanamurthy, ``An architecture for parallel topic models," *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, pp. 703--710, 2010.

[31] J. Varadarajan, R. Emonet, and J.-M. Odobez, ``Probabilistic latent sequential motifs: Discovering temporal activity patterns in video scenes," tech. rep., Idiap, 2010.

[32] L. L. Mølgaard, J. Larsen, and C. Goutte, ``Temporal analysis of text data using latent variable models," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing, 2009.*, 2009.

[33] D. M. Blei and J. D. Lafferty, ``Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning*, pp. 113--120, ACM, 2006.

[34] J. Chang and D. M. Blei, ``Relational topic models for document networks," in *International Conference on Artificial Intelligence and Statistics*, pp. 81--88, 2009.

[35] X. Tao, Y. Li, and N. Zhong, ``A personalized ontology model for web information gathering," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 23, no. 4, pp. 496--511, 2011.

[36] G. Liu, H. Shen, and L. Ward, ``An efficient and trustworthy p2p and social network integrated file sharing system," *Computers, IEEE Transactions on*, vol. 64, no. 1, pp. 54--70, 2015.

[37] M. Yang and Y. Yang, ``An efficient hybrid peer-to-peer system for distributed data sharing," *Computers, IEEE Transactions on*, vol. 59, no. 9, pp. 1158--1171, 2010.

[38] A. Iamnitchi, M. Ripeanu, E. Santos-Neto, and I. Foster, ``The small world of file sharing," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 22, no. 7, pp. 1120--1134, 2011.

[39] J. Tan, ``Discusses of user interest model in personalized search," *IJACT: International Journal of Advancements in Computing Technology*, vol. 5, no. 1, pp. 619--626, 2013.

[40] Y. Ye, G. Wu, and X. Luo, ``Research on interest model of user behavior," in *Proceedings of 2011 International Conference on Computer Science and Information Technology (ICCSIT 2011)*, 2011.

[41] S. Gong, ``The personalized information retrieval model based on user interest," *Physics Procedia*, vol. 24, pp. 817--821, 2012.

[42] A. McCallum, K. Nigam, *et al.*, ``A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752, pp. 41--48, Citeseer, 1998.

[43] A. K. McCallum, ``Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering." http://www.cs.cmu.edu/ mccallum/bow, 1996.

[44] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, ``A survey on visual content-based video indexing and retrieval," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 41, no. 6, pp. 797--819, 2011.

[45] D. Heckerman, *A tutorial on learning with Bayesian networks*. Springer, 1998.

[46] D. Inouye, P. Ravikumar, and I. Dhillon, ``Admixture of poisson mrfs: A topic model with word dependencies," in *Proceedings of The 31st International Conference on Machine Learning*, pp. 683--691, 2014.

[47] C. Sutton and A. McCallum, ``An introduction to conditional random fields," *Machine Learning*, vol. 4, no. 4, pp. 267--373, 2011.

[48] X. Yang, Y. Guo, and Y. Liu, ``Bayesian-inference-based recommendation in online social networks," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 24, no. 4, pp. 642--651, 2013.

[49] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra, ``Clustering on the unit hypersphere using von mises-fisher distributions," in *Journal of Machine Learning Research*, pp. 1345--1382, 2005.

[50] Y. Shi, M. Larson, and A. Hanjalic, ``Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges," *ACM Computing Surveys (CSUR)*, vol. 47, no. 1, p. 3, 2014.

[51] L. Yang, M. Qiu, S. Gottipati, F. Zhu, J. Jiang, H. Sun, and Z. Chen, ``Cqarank: jointly model topics and expertise in community question answering," in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pp. 99--108, ACM, 2013.

[52] S. Hingmire, S. Chougule, G. K. Palshikar, and S. Chakraborti, ``Document classification by topic labeling," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 877--880, ACM, 2013.

[53] Y. Chen, H. Amiri, Z. Li, and T.-S. Chua, ``Emerging topic detection for organizations from microblogs," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 43--52, ACM, 2013.

[54] J. H. Lau, ``Improving the utility of topic models: an uncut gem does not sparkle," 2013.

[55] J. Xia, F. Wu, C. Xie, and J. Tu, ``Inbi: An improved network-based inference recommendation algorithm," in *Networking, Architecture and Storage (NAS), 2012 IEEE 7th International Conference on*, pp. 99--103, IEEE, 2012.

[56] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, ``Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pp. 248--256, Association for Computational Linguistics, 2009.

[57] Y. Lv, T. Moon, P. Kolari, Z. Zheng, X. Wang, and Y. Chang, ``Learning to model relatedness for news recommendation," in *Proceedings of the 20th international conference on World wide web*, pp. 57--66, ACM, 2011.

[58] M. Bendersky and W. B. Croft, ``Modeling higher-order term dependencies in information retrieval using query hypergraphs," in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 941--950, ACM, 2012.

[59] K. Gimpel, ``Modeling topics," *Inform. Retrieval*, vol. 5, pp. 1--23, 2006.

[60] A. Andoni and P. Indyk, ``Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions," in *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pp. 459--468, IEEE, 2006.

[61] M. Girolami and A. Kabán, ``On an equivalence between plsi and lda," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 433--434, ACM, 2003.

[62] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, ``Optimizing semantic coherence in topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 262--272, Association for Computational Linguistics, 2011.

[63] D. M. Blei, ``Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77--84, 2012.

[64] S. Abbar, S. Amer-Yahia, P. Indyk, and S. Mahabadi, ``Real-time recommendation of diverse related articles," in *Proceedings of the 22nd international conference on World Wide Web*, pp. 1--12, International World Wide Web Conferences Steering Committee, 2013.

[65] M. Tavakolifard, J. A. Gulla, K. C. Almeroth, J. E. Ingvaldesn, G. Nygreen, and E. Berg, ``Tailored news in the palm of your hand: a multi-perspective transparent approach to news recommendation," in *Proceedings of the 22nd international conference on World Wide Web companion*, pp. 305--308, International World Wide Web Conferences Steering Committee, 2013.

[66] K. L. Caballero, J. Barajas, and R. Akella, ``The generalized dirichlet distribution in enhanced topic detection," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 773--782, ACM, 2012.

[67] Z. Chen and B. Liu, ``Topic modeling using topics from many domains, lifelong learning and big data," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 703--711, 2014.

[68] X. Wang and A. McCallum, ``Topics over time: a non-markov continuous-time model of topical trends," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 424--433, ACM, 2006.

个人简历、在读期间发表的学术论文与研究成果

个人简历：

先毅昆，男，1989 年 11 月出生。

2012 年 6 月毕业于同济大学软件学院，软件工程专业，获学士学位。

2012 年 9 月入同济大学软件学院，软件工程专业，攻读硕士学位。

已发表论文：

[1] Yikun Xian, Jie Huang, Yefim Shuf, Gene Fuh, and Zhen Gao. An Approach for In-Database Scoring of R Models on DB2 for z/OS. Rough Sets and Knowledge Technology. Springer International Publishing, 2014. 376-385.

[2] Yikun Xian, Jiangfeng Li, Chenxi Zhang, Zhenyu Liao. Video Highlight Shot Extraction with Time-Sync Comment. ACM MobiHoc 2015 - HotPOST '15.

待发表论文：

发明专利：

[1] 张晨曦, 先毅昆, 李江峰. 一种用户兴趣获取与传播的方法和装置: 中国, 201410494809.4[P]. 2015.01.07.