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/

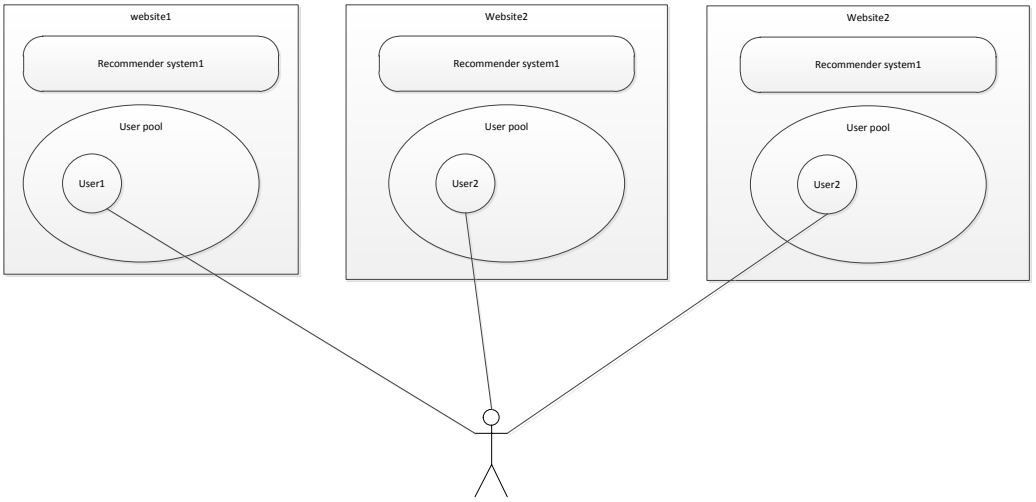

图 1.1: 互联网平台的信息孤岛现象

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

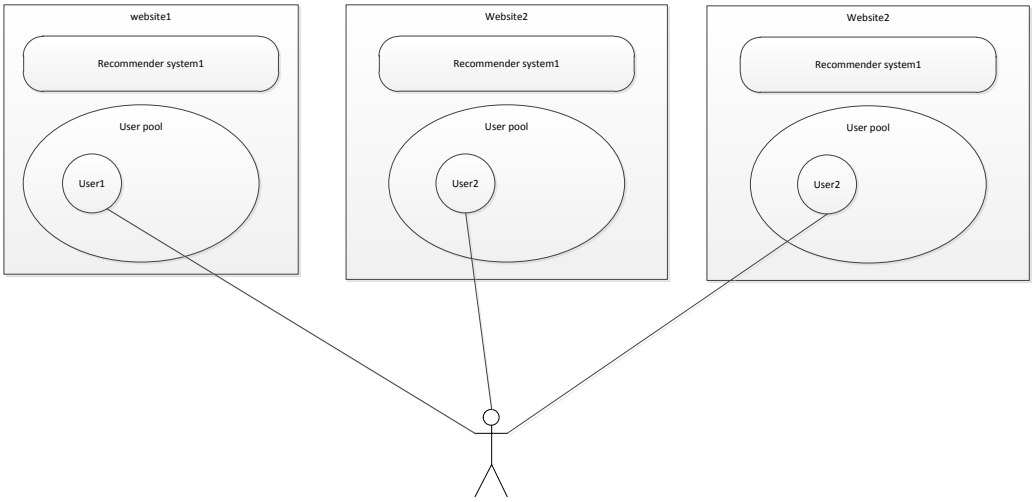

图 1.2: 基于 P2P 网络的智能系统架构

能够自主流动，就要解决这样几个问题：

- 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，例如普通的新闻或者博客，除开诸如作者、标题、时间等结构化信息，正文纯文本就是典型的非结构数据。因此
- 用户兴趣挖掘与关联：
- P2P 网络路由与搜索：

1.2 研究意义与应用价值

总结起来可以发现各个互联网平台之间相互封闭而导致的信息孤岛现象存在一下几个问题：

- 从计算资源的角度来看，在 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。从而使得硬件资源得到充分地利用，而不是像中心化架构中的服务器一样存在计算瓶颈。
- 从隐私安全的角度来看，所有的用户数据都存储在本地。
- 从拓扑架构的角度来看，扩展性好，。。。。

1.3 国内外研究现状

1.4 本文研究内容

1.5 本文内容安排

第 2 章　信息自主流动机制的方案概述

2.1　文本信息的匹配与推荐模型

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

第 3 章　互联网文本信息的匹配与推荐模型研究

第 4 章　用户兴趣模式的挖掘与描述模型的研究

第 5 章　基于 P2P 网络的信息自主流动机制的基础架构

第 6 章　实验分析与结果分析

第 7 章　原型系统设计与实现

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] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

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

[10] 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.

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

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

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

[27] 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.

[28] 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.

[29] 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.

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

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

[32] 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.

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

[34] 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.

[35] 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.

[36] 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.

[37] 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.