清 华 大 学

综 合 论 文 训 练

题目：大数据环境下信息抽取模板自动聚类与发现

系　　别：计算机科学与技术系

专　　业：计算机科学与技术

姓　　名：丘骏鹏

指导教师：朱小燕

辅导教师：郝宇

2013 年 6 月 5 日

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：学校有权保留学位论文的复印件，允许该论文被查阅和借阅；学校可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存该论文。

(涉密的学位论文在解密后应遵守此规定)

签　名：_____ 导师签名：_____ 日　期：_____

中文摘要

论文的摘要是对论文研究内容和成果的高度概括。摘要应对论文所研究的问题及其研究目的进行描述，对研究方法和过程进行简单介绍，对研究成果和所得结论进行概括。摘要应具有独立性和自明性，其内容应包含与论文全文同等量的主要信息。使读者即使不阅读全文，通过摘要就能了解论文的总体内容和主要成果。

论文摘要的书写应力求精确、简明。切忌写成对论文书写内容进行提要的形式，尤其要避免“第 1 章……；第 2 章……；……”这种或类似的陈述方式。

本文介绍清华大学论文模板 ThuThesis 的使用方法。本模板符合学校的本科、硕士、博士论文格式要求。

本文的创新点主要有：

- 用例子来解释模板的使用方法；
- 用废话来填充无关紧要的部分；
- 一边学习摸索一边编写新代码。

关键词是为了文献标引工作、用以表示全文主要内容信息的单词或术语。关键词不超过 5 个，每个关键词中间用分号分隔。（模板作者注：关键词分隔符不用考虑，模板会自动处理。英文关键词同理。）

关键词： TEX；LATEX；CJK；模板；论文

ABSTRACT

An abstract of a dissertation is a summary and extraction of research work and contributions. Included in an abstract should be description of research topic and research objective, brief introduction to methodology and research process, and summarization of conclusion and contributions of the research. An abstract should be characterized by independence and clarity and carry identical information with the dissertation. It should be such that the general idea and major contributions of the dissertation are conveyed without reading the dissertation.

An abstract should be concise and to the point. It is a misunderstanding to make an abstract an outline of the dissertation and words "the first chapter", "the second chapter" and the like should be avoided in the abstract.

Key words are terms used in a dissertation for indexing, reflecting core information of the dissertation. An abstract may contain a maximum of 5 key words, with semicolons used in between to separate one another.

Keywords: TEX; LATEX; CJK; template; thesis

目 录

第 1 章　引言

1.1　研究背景

1.1.1　大数据研究背景

随着互联网的快速发展，互联网上的信息呈现了爆发式的增长，互联网已经成为人们获取信息的一个主要渠道。然而，随着我们可以获得的数据量不断增加，人们的研究工作也受到了新的挑战。传统的数据处理手段正愈发显示出其局限性，如何有效对海量的数据进行处理，进而挖掘出我们所需要的内容逐渐成为一个重要的问题。在这个背景下，“大数据”迅速成为计算机科学领域非常受关注研究方向。

之前数据挖掘方面许多研究，更多地是关注如何在有限数据的情况下尽可能多地提取出准确的我们关心的信息。由于受到数据量的限制，很多数据中隐藏的模式和信息并不能被有效地发现。如今，随着我们可以获得的数据量的增加，数据不再是我们研究中的瓶颈，海量的数据带来的是数据的冗余性。

1.2　相关工作

1.3　本文重点难点和主要内容

1.4　本章总结

第 2 章　系统框架

2.1　整体框架介绍

2.2　网页过滤模块

2.3　网页聚类模块

2.4　模板生成模块

2.5　本章小结

第 3 章　基于后缀树的重复记录检测

3.1　后缀树简介

3.2　Ukkonen 后缀树构建算法

3.3　重复记录检测算法

3.4　本章小结

第 4 章　网页相似度计算与聚类

4.1　基于最长公共子串的网页距离计算

4.2　算法优化与改进

4.3　聚类算法实现

4.4　本章总结

第 5 章　模板生成和内容提取

5.1　模板生成

5.2　内容提取

5.3　本章总结

第 6 章　系统实现和实验结果

6.1　系统具体实现

6.2　实验数据和实验环境

6.3　实验结果

6.4　结果分析和存在的问题

第 7 章　工作总结和未来展望

7.1　工作总结

7.2　未来工作展望

插图索引

表格索引

参考文献

[1] Knuth D E. The TEX Book. 15th ed., Reading, MA: Addison-Wesley Publishing Company, 1989

[2] Goosens M, Mittelbach F, Samarin A. The LATEX Companion. Reading, MA: Addison-Wesley Publishing Company, 1994: 112–125

[3] Gröning P, Nilsson L, Ruffieux P, et al. Encyclopedia of Nanoscience and Nanotechnology, volume 1. American Scientific Publishers, 2004: 547–579

[4] Krasnogor N. Towards robust memetic algorithms. In: Hart W, Krasnogor N, Smith J, (eds.). Proceedings of Recent Advances in Memetic Algorithms. New York: Springer Berlin Heidelberg, 2004: 185–207

[5] 阎真. 沧浪之水. 人民文学出版社, 2001: 185–207

[6] 班固. 苏武传. 见: 郑在瀛, 汪超宏, 周文复, 编. 传记散文英华. 武汉: 湖北人民出版社, 1998: 65–69

[7] Chafik El Idrissi M, Roney A, Frigon C, et al. Measurements of total kinetic-energy released to the $N = 2$ dissociation limit of H_2 — evidence of the dissociation of very high vibrational Rydberg states of H_2 by doubly-excited states. Chemical Physics Letters, 1994, 224(10):260–266

[8] Mellinger A, Vidal C R, Jungen C. Laser reduced fluorescence study of the carbon-monoxide nd triplet Rydberg series-experimental results and multichannel quantum-defect analysis. J. Chem. Phys., 1996, 104(5):8913–8921

[9] Shell M. How to Use the IEEEtran LATEX Class. Journal of LATEX Class Files, 2002, 12(4):100–120

[10] 猪八戒. 论流体食物的持久保存 [D]. 北京: 广寒宫大学, 2005

[11] Jeyakumar A R. Metamori: A library for Incremental File Checkpointing[D]. Blacksburg: Virgina Tech, June 21, 2004

[12] 沙和尚. 论流沙河的综合治理 [D]. 北京: 清华大学, 2005

[13] Zadok E. FiST: A System for Stackable File System Code Generation[D]. USA: Computer Science Department, Columbia University, May, 2001

[14] IEEE Std 1363-2000. IEEE Standard Specifications for Public-Key Cryptography. New York: IEEE, 2000

[15] Kim S, Woo N, Yeom H Y, et al. Design and Implementation of Dynamic Process Management for Grid-enabled MPICH. Proceedings of the 10th European PVM/MPI Users' Group Conference, Venice, Italy, 2003

[16] Kocher C, Jaffe J, Jun B. Differential Power Analysis. In: Wiener M, (eds.). Proceedings of Advances in Cryptology (CRYPTO '99), volume 1666 of *Lecture Notes in Computer Science*. Springer-Verlag, 1999. 388–397

[17] Woo A, Bailey D, Yarrow M, et al. The NAS Parallel Benchmarks 2.0. Technical report, The Pennsylvania State University CiteSeer Archives, December 05, 1995. http://www.nasa.org/

[18] 贾宝玉, 林黛玉, 薛宝钗, 等. 论刘姥姥食量大如牛之现实意义. 红楼梦杂谈, 1800, 224:260–266

[19] 王重阳, 黄药师, 欧阳峰, 等. 武林高手从入门到精通. 第 *N* 次华山论剑, 西安, 中国, 2006

致　谢

衷心感谢导师 xxx 教授和物理系 xxx 副教授对本人的精心指导。他们的言传身教将使我终生受益。

在美国麻省理工学院化学系进行九个月的合作研究期间，承蒙 xxx 教授热心指导与帮助，不胜感激。感谢 xx 实验室主任 xx 教授，以及实验室全体老师和同学们的热情帮助和支持！本课题承蒙国家自然科学基金资助，特此致谢。

感谢 ThuThesis，它的存在让我的论文写作轻松自在了许多，让我的论文格式规整漂亮了许多。

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，独立进行研究工作所取得的成果。尽我所知，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享有著作权的内容。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。

签　名：_____ 日　期：_____

附录 A　书面翻译

文档结构相似性算法调研

摘要

这篇论文对文档结构相似性算法做了简明的调研，包括优化的树编辑距离算法和各种近似算法。这些近似算法包括简单加权标签相似度算法，文档结构的傅里叶变换，和将连续序列技术应用到结构相似度计算上。我们展示了三个令人惊奇的结果。第一，傅里叶变换的方法是所有近似算法中最不精确的一个，同时也是最慢的一个。第二，优化的树编辑距离的算法并不一定是最好的用来将不同网站的网页进行聚类的算法。第三，对于许多应用来说，最简单的结构的近似可能是最有用也是最有效率的机制。

A.1　简介

随着万维网上大量的文档的出现，自动地处理这些文档，将其应用于信息抽取，近似聚类和搜索的需求越来越大。在这个领域的主要工作主要集中在文档的内容上。然而，虽然万维网的继续发展和进化，越来越多的信息被放在结构化的富文本中，从 HTML 转换到 XML。这个结构化的信息是一个文档意义的重要体现。从文档中辨别出结构上“相似”，或者结构上互相“包含”的那些文档的方法是一个非常重要那些相关的相似文档关联起来的机制，而这些文档可能包含不同的文本内容，那些基于文本内容的相似度算法起不到这样的作用。

现在已有几个文档结构在其中起到关键作用的应用。目前的信息提取算法隐式或显式地依赖文档的结构化元素。结构化的信息能帮助我们将大量的从不同网站上获得的网页整理成一些可以大致可以比较的集合。这就使得软件能够将可以抽取出正确结果的集合和那些不能抽取有用信息的集合分离开来。

结构相似度是一个非常重要的话题，有非常多的算法可以计算任意两个文档结构之间的最小编辑距离。然而，由于这些算法复杂度非常高，通常都需要

$O(n^2)$ 或者更多的时间去计算距离，因此创造一些更快的，但是距离的计算精确度有些许损失的算法是有可能的。

我们在这篇论文的第二节展示了当前用于检测结构相似度的算法的概述。之后在第三节我们描述了一个新的基于连续序列的计算结构相似度的近似算法。在第四节我们在速度和精确度上对比了一下这些近似度算法和优化的树编辑距离算法。在第五节中我们用对不同算法特点的概括进行了总结。

A.2 当前研究状况

基于树的文档结构性相似度的研究已经有很长的历史了。早期基于树的变换检测来自 [1]，[2]。近期 Shasha [3]，[4]，[5] 和 Chawathe [6]，[7] 做了一些关于树到树变换的研究，主要集中在如何创建一个最小的脚本用来进行树之间的转换。在将一些基于结构的相似度计算修改成半结构的格式方面也有很多的尝试，包括 NiagraCQ [8]，Xdiff [9]，适用于 XML 的 Xyleme [10]，[11]，以及 AIDE [12]，[13] 和适用于 HTML 的 ChangeDetector™。

在学期间参加课题的研究成果

个人简历

xxxx 年 xx 月 xx 日出生于 xx 省 xx 县。

xxxx 年 9 月考入 xx 大学 xx 系 xx 专业，xxxx 年 7 月本科毕业并获得 xx 学士学位。

xxxx 年 9 月免试进入 xx 大学 xx 系攻读 xx 学位至今。

发表的学术论文

[1] Yang Y, Ren T L, Zhang L T, et al. Miniature microphone with silicon- based ferroelectric thin films. Integrated Ferroelectrics, 2003, 52:229-235. (SCI 收录, 检索号:758FZ.)

[2] 杨轶, 张宁欣, 任天令, 等. 硅基铁电微声学器件中薄膜残余应力的研究. 中国机械工程, 2005, 16(14):1289-1291. (EI 收录, 检索号:0534931 2907.)

[3] 杨轶, 张宁欣, 任天令, 等. 集成铁电器件中的关键工艺研究. 仪器仪表学报, 2003, 24(S4):192-193. (EI 源刊.)

[4] Yang Y, Ren T L, Zhu Y P, et al. PMUTs for handwriting recognition. In press. (已被 Integrated Ferroelectrics 录用. SCI 源刊.)

[5] Wu X M, Yang Y, Cai J, et al. Measurements of ferroelectric MEMS microphones. Integrated Ferroelectrics, 2005, 69:417-429. (SCI 收录, 检索号:896KM.)

[6] 贾泽, 杨轶, 陈兢, 等. 用于压电和电容微麦克风的体硅腐蚀相关研究. 压电与声光, 2006, 28(1):117-119. (EI 收录, 检索号:06129773469.)

[7] 伍晓明, 杨轶, 张宁欣, 等. 基于 MEMS 技术的集成铁电硅微麦克风. 中国集成电路, 2003, 53:59-61.

研究成果

[1] 任天令, 杨轶, 朱一平, 等. 硅基铁电微声学传感器畴极化区域控制和电极连

接的方法: 中国, CN1602118A. (中国专利公开号.)

[2] Ren T L, Yang Y, Zhu Y P, et al. Piezoelectric micro acoustic sensor based on ferroelectric materials: USA, No.11/215, 102. (美国发明专利申请号.)