学校编码：10384　　分类号　　密级
学号：200428029　　UDC

硕 士 学 位 论 文

基于语义距离的文本聚类算法研究

Text Clustering Research Based On Semantic Distance

林 丽

指导教师姓名：冯 少 荣 副教授

专 业 名 称：计算机应用技术

论文提交日期：2007 年 4 月

论文答辩时间：2007 年 6 月

学位授予日期：

答辩委员会主席：

评 阅 人：

2007 年 4 月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文产生的权利和责任。

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1. 保密（），在年解密后适用本授权书。

2.不保密（　）

（请在以上相应括号内打“√”）

作者签名：　　日期：　　年　　月　　日

导师签名：　　日期：　　年　　月　　日

摘　要

网络技术迅速发展的今天，人们越来越感受到了信息的冲击，而文本是信息的重要载体，人们日常生活中所接触到的信息有80%左右以文本的形式存在。信息内容和格式的多样化、复杂化，使人们无法遍历所有感兴趣的内容，而且又不存在标准的文本分类准则，所以管理收集到的文本信息成为亟待解决的问题，对于文本聚类技术的研究更显重要。

现有的文本聚类方法大多采用基于 VSM 的关键词匹配来计算文本间相似度，这种方法的最大的缺点就是忽略了词之间的语义信息，忽略了各维度之间的联系，导至文本的相似度计算不够精确，所以本文从语义上具体分析文档，利用文本具体语义计算文本间的相似度，使得文本聚类结果更合理，主要工作及创新点有：

1、以《知网》作为语义的本体，利用语义距离计算文档间相似度，把文档间相似度计算具体转化为词语间语义距离、义原间语义距离。考虑到文本聚类具体应用，本文根据《知网》描述各个词的规律，改进现有词语相似度计算方法，更有利于发现词语的相关性，适应了文本聚类的要求。

2、文本聚类算法主要采用一次遍历聚类算法即最近邻聚类算法，并提出第二次聚类方法改进最近邻算法对输入次序敏感的问题。类中心方面，引入相似权重的概念，并根据权重优胜略汰候选类特征词，使得最后选择的类特征词能够代表类的主题，达到文本聚类的目的。

论文最后实验语料来源于中科院的中文自然语言处理开放平台(CNLP)网站，下载了100篇文档对所提出的算法进行了实验，并利用聚类精度和召回率对实验结果进行了评价，然后把评价结果与基于VSM的K-Means聚类算法进行了比较，结果证明本文所提出的基于语义距离文档聚类算法在聚类精度和召回率上都优于基于VSM的K-Means聚类算法，达到了算法改进的目的。另外基于语义距离的文档聚类结果显示它还能从语义上更加细分主题，为用户收集文本信息提供更好的导航。

关键词：文本聚类；语义距离；《知网》；VSM；K-Means；最近邻聚类

Abstract

Today, as the rapid development of network, people have a growing feeling about the information impact. Text is the important carriers of information 80% of the daily information people have touched is in the form of text. The information's content and format are so various and complicated that people are unable to traverse all their interested, but there is still no standard criteria for the classification of text, so it needs urgent solution to manage the collected information from the text. As a result the research of the text clustering technology is more important.

Most of the current clustering methods use keyword matching based on VSM to calculate text similarity. The major drawback of this approach is it overlooks semantic information between words and the link between the various dimensions, and the result of the text similarity isn't accurate. So the paper analysis the text from the semantic, use the specific semantic of the text to compute the text similarity, the test proves the result is more reasonable. The major contributions are as follows:

1、We use the famous Chinese knowledge library-《Hownet》to calculate the similarity between documents, the calculation is decomposed to several parts including semantic distance between keywords and between atoms. Considering the specific application of the text clustering, the paper uses the rules which 《Hownet》 describe the words to improve the existing words similarity calculation, this improvement can find the relevance between words and fit the requirements of the text better.

2、Our clustering algorithm mainly uses single pass clustering (nearest neighbor clustering),and proposes the second clustering to improve the weakness of nearest neighbor clustering which is sensitive to the input order of the document. In respect of category center, the similar weight concept is introduced, we choose some feature words to represent the cluster according the weight, the remaining feature words last are similar with the main themes of the cluster, achieve the purpose of text clustering.

Finally, the proposed algorithm is implemented and the testing experiments are conducted with 100 documents downloaded from CNLP Platform. Using the precision

and recall of clustering as the evaluation of result, we compare the clustering results of the proposed algorithm with the K-Means algorithm base on VSM, the experiments indicated that the performance of the proposed algorithm is better than the VSM+K-Means algorithm. Moreover, the text clustering based on semantic distance shows it can divide the main theme into sub-themes, and these sub-themes can provide better navigation for the information collection.

Key Words: Text Clustering; Semantic Distance;《Hownet》;VSM; K-Means; Single Pass Clustering

目　录

Contents

第一章 绪论

1.1 研究背景

当今 Internet 已经成为人们获取信息的重要来源.Web2.0 的出现使 Internet 和人类的关系变得越发紧密。但是面对大量的电子信息成几何级数增长，人们从大规模的文本中快速获取所需要的信息的要求，日益变得迫切，而基于数据挖掘的机器学习技术为在大量原始数据中提取信息提供了方法。机器学习主要分为两个分支：有监督的学习和无监督的学习，而在无监督学习领域，聚类技术是主要的无监督学习的工具，聚类技术与文本挖掘技术的结合产生了文本聚类技术。

文本聚类主要是依据著名的聚类假设：同类的文档相似度较大，而不同类的文档相似度较小。作为一种无监督的机器学习方法，聚类由于不需要训练过程，以及不需要预先对文档手工标注类别，因此具有一定的灵活性和较高的自动化处理能力，已经成为对文本信息进行有效地组织、摘要和导航的重要手段，为越来越多的研究人员所关注。

文本聚类主要应用有：

- 文档聚类可以作为多文档自动文摘等自然语言处理应用的预处理步骤，如将每天发生的重要新闻文本进行聚类处理，并对同主题文档进行冗余消除、信息融合、文本生成等处理，从而生成一篇简明扼要的摘要文档；
- 对搜索引擎返回的结果进行聚类，使用户迅速定位到所需要的信息。Hua-Jun Zeng等人提出了对搜索引擎返回的结果进行聚类的学习算法。比较典型的系统则有vivisimo（http://www.vivisimo.com）和infonetware（http://www.infonetware.com）等。系统允许用户输入检索关键词，而后对检索到的文档进行聚类处理，并输出各个不同类别的简要描述，从而可以缩小检索的范围，用户只需关注比较有希望的主题。另外这种方法也可以为用户二次检索提供线索；
- 对用户感兴趣的文档（如用户浏览器cache中的网页）聚类，从而发现用户的兴趣模式并用于信息过滤和信息主动推荐等服务。

- 数字图书馆服务。通过SOM神经网络等方法，可以将高维空间的文档拓扑保序地映射到二维空间，使得聚类结果可视化和便于理解。
- 文档集合的自动整理。

1.2 文本聚类算法综述

聚类算法的研究早在 20 世纪 60 年代就开始了,但是受当时各方面条件的限制并没有太大的发展,直到 20 世纪 90 年代才引起了广泛的关注,并取得非常重大的突破,包括 K-Means、CLAR-ANS、EM、BIRTH、DBSCAN、STING、CLIQUE 等在内的一大批新的聚类算法被陆续提出,目前仍旧是一个非常热点的问题. 但是,这些算法中大多数不是针对文本数据而提出的,只有少部分能被应用于文本聚类。大体上,文本聚类算法大致分为划分聚类算法、层次聚类算法、基于密度的聚类算法、基于模型的聚类算法和基于网格的聚类算法、STC 算法以及其他算法等几大类。

1、划分聚类方法。

给定 n 个对象，一个划分方法构建对象的 k 个划分，每个划分表示一个聚簇，并且 k<n。给定要构建的划分的数目 k，划分方法首先创建一个初始划分。然后采用一种迭代的重定位技术，尝试通过对象在划分间移动来改进划分。一个好的划分的一般准则是：在同一个类中的对象之间尽可能“接近”或相关，而不同类中的对象之间尽可能“远离”或不同。目前比较流行的是 k-平均算法，k 一中心点算法两种启发式的划分方法，尤其是 K-平均算法常常在文本聚类中使用。这类算法虽然运行时间快但是存在一些缺点如受初始簇中心影响较大，要预先指定聚类数 k，容易受孤立点的影响等。

2、层次聚类方法

层次聚类算法将文本数据对象组成一棵聚类的树。根据层次的分解是自底向上还是自顶向下形成,层次聚类算法又进一步分为凝聚的和分裂的层次聚类算法. 在凝聚的层次聚类算法中,开始时每一个文本对象都作为单独的一个组,然后相继地合并最相近的两个组,直到只剩下一个组或者达到一个终止条件。分裂的层次聚类算法与凝聚的层次聚类算法刚好相反,它开始时将所有的文本对象都放在一个组中,然后在每一步中,将一个组分裂为两个更小的组,直到所有的文本对象都

各自成组。层次聚类有一个更大的缺点在于其不可逆性,也就是说一旦一次合并或者分裂完成,就不能被撤销，最典型的算法就是 Single-Link 算法.这类算法虽然能够分层展示文本数据，但是时间复杂度为 $O(n^2)$,而且容易产生链式聚类。

3、基于密度的聚类算法

基于密度的聚类算法最大的优点就在于能够发现任意形状的簇,之所以具有这样的优点是因为它将簇看成是数据空间中被低密度区域分割开的高密度对象区域。采用如下策略：只要临近区域的密度(对象的数目)超过某个阈值,就继续聚类。DBSCAN 是最具代表性的基于密度的聚类算法,它也被常用于文本聚类。其基本思想是:对于一个簇中的每一文本对象,在其给定半径(用 ε 表示)的领域中包含的文本对象数目不小于某一给定的最小数目(用 MinPts 表示).这类算法虽然它能发现任意形状的簇，但对参数 ε 和 MinPts 敏感

4、基于模型的聚类算法

基于模型的聚类算法试图优化给定的数据和某些数学模型之间的适应性.。这样的方法经常是基于这样的假设：为每一个簇假定了一个模型,然后寻找数据与给定模型的最佳拟合。这类算法主要分为两类:统计学方法和神经网络方法,自组织特征映射(self - organizing feature map, SOM)是一种利用了人工神经网络技术的聚类方法[12]。

SOM 是 Kohonen 于 1981 年提出的,这种网络模拟大脑神经系统的功能,是一种竞争式学习网络,在学习中能无监督地进行自组织学习。SOM 网络结构是由输入层和竞争层组成。输入层神经元数为 n,竞争层由 M=m2 个神经元组成的二维平面阵列,输入层与竞争层各神经元之间实现全互连接。

SOM 网络的工作原理是将任意维输入模式在输出层映射成一维或二维离散图形，并保持其拓扑结构不变。此外,网络通过对输入模式的反复学习，可以使权重向量空间与输入模式的概率分布趋于一致，即权重向量空间能反映输入模式的统计特征，这种自组织聚类过程是系统自主、无导师指导的条件下完成的。SOM 聚类算法具有无需监督，能自动对输入模式进行聚类的优点，但是也存在一些问题：在数据量大情况下，随着学习次数增多，SOM 网络的学习效果反而降低，又称学习过度。另外 SOM 要求输出神经元数很大，因而其权重向量数目也很多，这使它的规模太大。

5、基于网格的聚类算法

基于网格的聚类算法把对象空间量化为有限数目的单元，形成了一个网络结构。所有的聚类操作都在这个网格结构(即量化的空间)上进行。本质上，它经过了如下的转换过程：数据一>网格数据一>空间分割一>数据分割。这样不直接对数据进行处理的优点是网格数据的增加使得基于网格的聚类技术不受数据次序的影响，因此处理速度非常快，其处理时间独立于数据对象的数目，只与量化空间中每一维的单元数目有关。STING 是基于网格聚类算法的典型例子，CLIQUE 和 WaveCluster 这两种算法既是基于网格的,又是基于密度的聚类算法。

6、后缀树（STC）聚类算法

后缀树聚类算法:后缀树聚类算法是一种直观的文本聚类算法，它将文本聚类为一组的依据是文本含有共同的短语。实际上是将文本看成词的序列，充分利用了词与词之间的距离信息，在寻找文本共同的短语的过程中使用了后缀树这种数据结构。Oren Zamir 首先提出了后缀树聚类算法,并且将这种技术运用到搜索结果的可视化中，取得了很好的效果，具体的算法过程可以参考文献[2]。搜索引擎 Carrot2 中的部分工作也利用了 STC 算法实现 Web 搜索结果的聚类，并且指出了 STC 算法的若干不足,但没有对聚类结果的质量进行评价[3]。

除了上述常用的聚类分析算法外，Web 文本聚类分析还采用基于群体智能的 Web 文本聚类算法、基于关联规则的 Web 文档聚类算法、简单贝叶斯聚类法及基于概念的文本聚类算法等。

1.3 本文主要工作

上述文本聚类算法大部分是基于 VSM（文本矢量）模型来计算文本间相似度，这种计算方法假设词语间是独立的，并没有从语义上去分析文档内容，因而不能准确计算文档间的相似度，影响了聚类的精度。

本文着重从语义上分析文本，先根据 TF*IDF 选择文档的关键词，把文档表示为一组关键词集合，再利用语义距离计算关键词间的相似度，通过相似关键词对数目最后衡量文本间的相似度，这样使得文本之间的相似值真正从语义上具体分析，结果更接近人主观衡量。

在计算词语相似度方面，考虑到文本聚类的具体应用，本文在文献[18]的计