硕 士 学 位 论 文

高校毕业生就业推荐系统的设计与开发

The Design and Implement of Graduate Occupation Recommending System

作 者 姓 名：吴 迪

学科、 专业：计算机应用技术

学　　号：20809429

指 导 教 师：林鸿飞

完 成 日 期：2010.11.14

大连理工大学

Dalian University of Technology

大连理工大学学位论文独创性声明

作者郑重声明：所呈交的学位论文，是本人在导师的指导下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内容和致谢的地方外，本论文不包含其他个人或集体已经发表的研究成果，也不包含其他已申请学位或其他用途使用过的成果。与我一同工作的同志对本研究所做的贡献均已在论文中做了明确的说明并表示了谢意。

若有不实之处，本人愿意承担相关法律责任。

学位论文题目：高校毕业生就业推荐系统的设计与开发

作 者 签 名 ：_____ 日期：____年____月____日

摘　　要

近年，随着高校毕业生数量持续增长和全球金融海啸给我国经济带来的不利影响，高校毕业生的就业形势日趋严峻。而当前我国各所高校的毕业生就业工作尚不足以为每一名毕业生提供准确有效的就业指导和就业推荐，各高校的就业网更多是仅提供招聘信息发布功能，并不具备信息推荐功能。“高校毕业生就业推荐系统”的设计开发则刚好填补了这份空白。通过“高校毕业生就业推荐系统”，毕业生可以根据自己的个体情况，得到一份科学、可靠的就业推荐，并以此作为择业依据。

针对现有网络求职平台在就业推荐过程中存在的缺陷，同时结合高校毕业生求职和企业校园招聘的特点，我们设计了“高校毕业生就业推荐系统”。在系统的设计过程中，我们通过比较应届和往届毕业生基本特征，分别采用基于经验公式和基于 SimRank 算法两种办法来获得两名学生之间的相似度。随后，根据学生之间的相似度，通过 K-Means 算法对学生进行聚类分析，并通过进一步分析得到应届毕业生与企业间的相似度。最后，本文将学生与企业的相似度同基于 PageRank 算法获得的各个企业的“求职指数”结合，从而获得企业的推荐排序权值，并根据这个权值将排序靠前的企业推荐给对应的应届毕业生。

尽管本文采用了两种不同的学生相似度计算方法，但通过本文第五章的测试对比实验，在最终系统中，我们选择基于经验公式计算学生间相似度的方法来完成学生间相似度计算。根据测试实验的结果，我们认为：本系统不仅功能上符合设计初衷，能够有效的为毕业生提供就业推荐服务，推荐结果科学合理；同时还能够帮助缺乏求职目标的学生制定求职目标，提升学生的求职成功率，在一定程度上降低学生求职成本。对比之前就业网单纯的信息发布功能，本系统提供的就业推荐功能具有较高的实际应用价值。

关键词：推荐系统；SimRank；PageRank；聚类分析；K-Means

The Design and Implement of Graduate Occupation Recommending System

Abstract

In recent years, because of the continuously growing number of the college graduates and the adverse impact on China's economy caused by the global financial tsunami, the college graduates faced an increasingly tough job market. However, at present, the graduates-employment service of all the universities in China cannot fully provide each graduate with proper and effective employment guidance and job recommending. Besides, what universities' employment networks have is only recruitment information releasing function, but not information recommending function. The design and development of "*the employment recommending system for the college graduates*" can just fill the gap. By using this system, the college graduates can get a scientific and reliable employment recommendation according to their individual circumstances. With the recommendation, the graduates can make a wiser choice for their career.

As the existing network platform in the employment recommending process has flaws, we have designed "*the employment recommending system for the college graduates*", *which has taken the* features of graduates' seeking jobs as well as campus recruitment into consideration. In the system design process, we compare the basic features of fresh and previous graduates and get the similarity of these two groups of graduates by using the empirical formula and Simrank algorithm respectively. Then, based on the result, we obtain the similarity of the fresh graduates and enterprises in further cluster analysis. Finally, we get all enterprises' recommendation-ranking weight by combining the similarity (of the fresh graduates and enterprises) with the enterprise's "Job Index" which is obtained by PageRank algorithm. Thus, several top enterprises in our rankings will be recommended to the graduates.

Though this article has applied two different algorithms for calculating the similarity between graduates, we choose the empirical formula in the final system according to the result from the comparative testing experiment in Chapter V of this article. We can conclude from the testing experiment result that the final system not only meet the original intention in function, which can effectively offer scientific and reasonable employment recommendation to the graduates, but also help those who lack job objectives to set one and enhance the success rate in seeking a job, which to some extent means reducing the cost in job seeking. Compared with the

simple information releasing function of the present employment network, the employment recommendation function in our system has a higher practical value.

Key Words：Recommending System;SimRank;PageRank;Clustering Analysis;K-Means

目　　录

1 绪论

1.1 大学生职业推荐系统的开发背景及意义

近年来随着高校毕业生数量连年增长和全球金融海啸给我国经济带来的不利影响，高校毕业生的就业形势日趋严峻。然而，我国各所高校的毕业生就业工作目前尚不足以为每一名毕业生提供准确有效的就业指导和就业推荐，绝大部分高校的网络就业平台仅提供企业招聘信息发布的功能，无法向毕业生提供信息推荐的功能，如我校目前正在使用的就业网站（如图 1.1 所示）虽包含了诸如在线简历投递、职业测评等多项功能，但仍然无法实现针对每一名毕业生推荐不同企业招聘信息的功能[1]。

图 1.1 大连理工大学就业网

Fig. 1.1 Employment Website of DUT

本文基于对大连理工大学毕业生就业市场的调查，发现：针对某所特定高校，其就业市场一般相对稳定，企业需求相对固定，学生签约单位范围每年不会出现太大的变化。即使相同专业不同学历的毕业生就业市场会有一定区别，但各自均拥有其相对稳定的就业市场。通过对招聘企业和应、往届毕业生的调查，我们得出结论：大连理工大学应届毕业生在求职应聘过程中参考往届学生就业去向的比例超过 87%，企业根据以往的招聘经验，也更愿意招聘那些与往年录用学生比较类似的应届毕业生（如专业背景、学历、

生源地、政治面貌、学生工作经历、外语水平等方面）。

同时，根据调查发现：有超过 90%的往届毕业生是通过学校的就业网站获得到自己签约企业的招聘信息，由此可见，学校的就业网在学生整个求职应聘过程中的重要作用。每当大学生求职旺季，学校就业网每天发布的招聘信息达到 30-50 条，每年累计发布的招聘信息将达到 3000 条以上,甚至在每年春秋两季大型毕业生双选会期间，就业网一天内发布的企业招聘信息超过 400 条。而目前，学校的就业网仅仅具有简单的信息筛选功能，并不具备招聘信息推荐功能，毕业生要在数量庞大的招聘信息中寻找适合自己的企业招聘信息，工作量较大，耗时较多。针对此问题，我们提出了“高校毕业生就业推荐系统”的设计构想。通过“高校毕业生就业推荐系统”，每一名应届毕业生可以根据自己的个体情况，得到一份科学、可靠的就业推荐，在获取招聘信息的过程中减少浏览不相关招聘信息所损失的时间，并可以以我们的推荐作为自己择业的参考依据。

目前，“高校毕业生就业推荐系统”已经通过了前期测试，从测试者反馈信息看，系统的信息推荐准确率较高，符合设计要求；并且，本系统操作简便，推荐企业排序合理，界面表达直观，能够满足用户的需要。

1.2 目前国内相关职业推荐平台的调研分析

通过调查我们发现，目前国内功能相对完善的求职平台包括前程无忧网（http://www.51job.com/）、智联招聘网（http://www.zhaopin.com/）等。上述平台在求职推荐功能上，具有如下特点：

（1）前程无忧网的职位推荐功能主要基于用户以往投递简历的企业类别、薪酬水平、从事岗位、企业地域等信息对用户进行相似职位的推荐，同时在推荐方法上采用了协调过滤技术为新用户进行推荐；

（2）智联招聘网的职业推荐功能与前程无忧网的职业推荐具有一定区别，一方面对于新用户根据用户的基本信息，如学历、求职目标、目标行业类别等将职位信息分类，另一方面也会根据用户以往的职位选择进行推荐。

（3）在无简历或者尚未投递简历的情况下，前程无忧网还提供另一种职位推荐功能（如图 1.2 所示），但该功能尚不完善，目前仅仅能够根据用户的求职目标和工作地点进行推荐；

通过对其他求职类网站的调研后，我们发现：绝大部分求职类平台提供的职位推荐服务是基于协同过滤技术和分类的方法。但是，由于无法获得用户最终是否被推荐职位录用的信息（客观事实对推荐结果的评价），这些推荐系统都无法为用户提供高质量，

高准确率的求职信息推荐功能。

图 1.2　前程无忧网职位推荐模块

Fig. 1.2　Module For Recommendation in 51Job

1.3　本文主要工作

针对现有网络求职平台在就业推荐过程中存在的缺陷，同时结合高校毕业生求职和企业校园招聘的特点，我们设计了“高校毕业生就业推荐系统”。在系统的设计过程中，我们提出了基于 SimRank 算法计算学生相似度的方法，同时，我们也采取了通过经验公式计算学生相似度的方法，通过两种方法的对比测试，我们发现两种方法均能满足系统应用需要，但采用基于 SimRank 算法计算学生相似度的方法系统最终推荐准确率较高。

与此同时，在“高校毕业生就业推荐系统”的开发过程中，我们还引入了“企业求职指数”的概念，并设计了基于 PageRank 算法的企业求职指数计算方法。

最后，我们通过进行多组测试对比实验对系统的参数选取和应用价值进行了进一步的分析。

1.4　本文的篇章结构

本文叙述了“高校毕业生就业推荐系统”的设计开发过程。组织如下：第二章相关技术背景，主要包括关于系统中所应用的各项技术的相关介绍；第三章系统概述，主要包括系统的模块设计等内容；第四章关键技术，结合系统模块设计着重介绍各模块应用的相关技术和算法；第五章测试分析及系统评价，结合对比测试结果和功能测试反馈，对系统的应用价值进行进一步探讨。

2 相关技术背景

2.1 几种推荐方法的介绍

一般的推荐系统普遍采用的框架结构如图 2.1 所示，从图中我们可以看出：推荐方法是整个推荐系统中最为核心、最为关键的部分，系统所采用的推荐推荐方法可以在很大程度上决定这个推荐系统性能的优劣。目前，比较主流的推荐方法包括：协同过滤推荐、基于内容推荐、基于效用推荐、基于关联规则推荐、组合推荐和基于知识推荐等[2-4]。

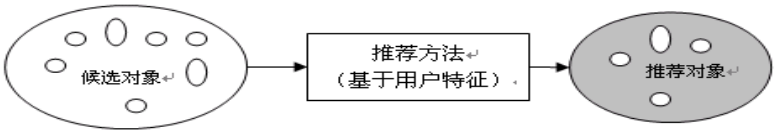

图 2.1 一般推荐系统的框架图

Fig. 2.1 General Framework of Recommendation System

对于上述各种推荐技术，通过对比我们发现：每种推荐方法都有其各自的优点和缺点（见表 2.1）。

表 2.1 各种推荐方法的比较

Tab. 2.1 The Comparison of All Recommendation Methods

推荐方法	优点	缺点
基于内容推荐	1. 推荐结果直观，容易解释； 2. 不需要领域知识	1. 稀疏问题；新用户问题； 2. 复杂属性不好处理； 3. 要有足够数据构造分类器
协同过滤推荐	1. 新异兴趣发现、不需要领域知识； 2. 随着时间推移性能提高； 3. 推荐个性化、自动化程度高； 4. 能处理复杂的非结构化对象	1. 稀疏问题； 2. 可扩展性问题； 3. 新用户问题； 4. 质量取决于历史数据集； 5. 系统开始时推荐质量差；
基于规则推荐	1. 能发现新兴趣点； 2. 不要领域知识	1. 规则抽取难、耗时； 2. 产品名同义性问题； 3. 个性化程度低；
基于效用推荐	1. 无冷开始和稀疏问题； 2. 对用户偏好变化敏感； 3. 能考虑非产品特性	1. 用户必须输入效用函数； 2. 推荐是静态的，灵活性差； 3. 属性重叠问题；
基于知识推荐	1. 能把用户需求映射到产品上； 2. 能考虑非产品属性	1. 知识难获得； 2. 推荐是静态的

2.2 SimRank 算法介绍

SimRank 算法的核心思想是：任意 2 个对象是相似的，并且与它们相关联的对象之间也是相似的，在这样的情况下，SimRank 算法可以做为计算任意 2 个对象之间相似度的一种方法[5,6]。

为便于 SimRank 算法的理解，我们利用有向图的方式来表示任意对象之间的关联，通过在关联的有向图上进行迭代计算来获得任意点（对象）之间的相似度。在迭代开始时任意 2 个不同对象之间的相似度都初始化为 0，任意对象与其自身的相似度都初始化为 1(最高相似度)。图 2.2 给出了一个 SimRank 迭代所使用的关联有向图示例（图 2.2），该图表示不同网页之间的链接关系，包括学校的网页(Univ)，2 个教授的网页(ProfA 和 ProfB)和 2 个学生的网页(StudentA 和 StudentB)。因此 ProfA 和 ProfB 都与 Univ 关联，所以可以认为 ProfA 和 ProfB 具有相似关系。同理，StudentA 和 StudentB 分别与 ProfA 和 ProfB 关联，由于 ProfA 和 ProfB 相似，因此可以推断出 StudentA 和 StudentB 在某种程度上也相似[7-10]。

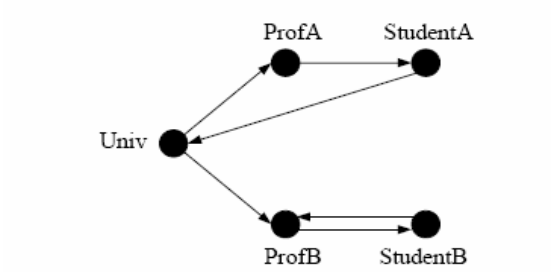

图 2.2 SimRank 迭代所使用的关联有向图示例

Fig. 2.2 Demonstration of Related Directed Graph under SimRankIteration

当然，在图 2.2 中，由于 StudentA 和 StudentB 仅仅能对应唯一的一个教授（ProfA 或 ProfB）所以图 2.2 也可以认为是 SimRank 算法的一种特殊情况。如果我们对 SimRank 算法的核心思想进行扩展，我们会得到图 2.3 所示的无向图示例。因图 2.3 中的学校（Univ）与专业（Major）之间的包含与归属两种关系，可以认为他们是一种双向的关联，这样，我们便可以用无向的边来表示这种关联。由于 UnivA 与 UnivB 都与 MajorA 关联，所以可以认为 UnivA 与 UnivB 具有相似关系。同理，我们也可以认为 UnivA 与 UnivC 具有相似关系，并推断出 UnivB 与 UnivC 在某种程度上也相似。

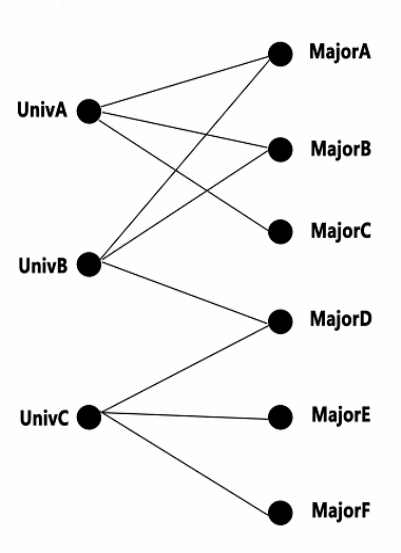

图 2.3　SimRank 迭代所使用的关联无向图示例

Fig. 2.3　Demonstration of Related Undirected Graph under SimRankIteration

在类似于图 2.3 的无向图中，对于图上任意一个节点 m，如果存在边<m, n>，则称 n 为 m 的邻居。假设 $I(m)$ 表示 u 的所有邻居的集合，$I_i(m)$ 表示 m 的第 i 个邻居，则有 $1 \le i \le I_i(u)$。

若定义 $S(a,b)$ 为对象 a 和 b 之间的相似度，$S(a,b) \in [0,1]$，则根据 SimRank 算法可定义 $S(a,b)$ 如下：

$$S(a,b)=\begin{cases} 1 & \text{若} a=b \\ \dfrac{C}{|I(a)||I(b)|}\sum_{i=1}^{|I(a)|}\sum_{j=1}^{|I(b)|}S(I_i(a),I_j(b)) & \text{若a} \neq \text{b} \end{cases} \qquad (2.1)$$

在公式 2.1 中 C 为衰减系数，是一个大于 0 且小于 1 的常数，如果 a 与 b 并不是同一个对象，并且 a 或者 b 没有邻居，则置 $S(a,b)=0$。显然，公式 2.1 所定义的 2 个对象之间的相似度，完全依赖于与这 2 个对象相关联的对象之间的相似情况。

在 SimRank 算法的运算过程中，通过按照公式 2.1 计算图中任意两点间的相似度，我们得到的将是一个方程组，方程组的元数取决于图中节点的数量，而方程中的未知量即是任意两点间的相似度。这样，两点间相似度的计算问题演变成为方程组的求解问题。

但是，往往在实际应用过程中，我们需要计算的对象集合会非常庞大，这样再通过简单的方程组求解，会变得非常的困难。这时，就需要通过迭代方法实现对象之间的相

似度计算的过程。首先，在迭代开始时，任意 2 个对象间的相似度按照下式（公式 2.2）初始化：

$$R_0(a,b)=\begin{cases}0 & \text{a}\neq\text{b}\\ 1 & \text{a}=\text{b}\end{cases} \tag{2.2}$$

公式 2.2 中，$R_0(a,b)$ 表示第 0 次迭代（初始状态）下 a 和 b 之间的相似度，以后每次迭代都在前一次已有结果的基础上，对相似度按照下式（公式 3）进行更新：

$$\text{R}_{k+1}(a,b)=\begin{cases}1 & \text{若}a=b\\ \dfrac{C}{|I(a)||I(b)|}\sum_{i=1}^{|I(a)|}\sum_{j=1}^{|I(b)|}R_k(I_i(a),I_j(b)) & \text{若a}\neq\text{b}\end{cases} \tag{2.3}$$

上述迭代过程不断进行，直到收敛为止。最后得到的 $R_{k+1}(a,b)$ 就是对象 a 和 b 之间的相似度 $S(a,b)=0$ 。

2.3 聚类分析介绍

在实际生活中，有大量的分类问题存在。“类”，简单来说，就是指相似元素所构成的集合。聚类分析（Clustering Analysis）是研究分类问题的一种统计分析方法[11]。

一个数据集合，通过聚类分析，可以分成若干个类，在这些类中，同一个类内的数据对象具有较高的相似度；而不同类内的数据我们可以认为是不相似或相似程度很低的。对于数据相似或不相似的区分，是通过某个数据的特征（属性）来确定的。通常，这种相似关系就是利用各个数据之间的相似度（距离）来进行表述[12]。

根据聚类分析的特点，其被广泛应用于数据挖掘、统计学和机器学习等多个方向，其典型应用主要包括：在商业领域，通过聚类分析，市场人员可以快速的发现顾客群中所存在的具有不同特征的群体，并可以利用购买模式来描述这些全体，进而针对不同的群体制定不同的销售计划；在生物领域，通过聚类分析，科研人员可以很容易获取动物或植物所存在的层次结构，进而加深对不同的动物或植物物种的认识。此外，聚类分析还被广泛应用于识别和分类互联网上的各种文档以便进行更好的对这些文档进行数据挖掘[13]。

作为数据挖掘的一项功能，聚类分析还可以作为一个单独使用的工具，来帮助分析数据的分布、了解各数据类的特征、确定所感兴趣的数据类以便作进一步分析。当然聚类分析也可以作为其它算法（诸如：分类和定性归纳算法）的预处理步骤。数据聚类分析是一个正在蓬勃发展的领域。聚类分析所涉及的领域包括：数据挖掘、统计学、机器

学习、空间数据库技术、生物学和市场学等。由于各应用数据库所包含的数据量越来越大，聚类分析已成为数据挖掘研究中一个非常活跃的研究课题。目前广泛使用的聚类分析方法主要有如下几种[14]：

①划分法(partitioning methods)：对于给定的一个具有 N 个纪录的数据集，首先，通过对其构造 K 分组（K<N），我们可以认为每一个分组就代表一个类。并且这 K 个分组需要满足如下条件：（1）每一个分组至少包含一条纪录；（2）每一个纪录属于且仅属于一个分组（在某些模糊聚类算法中可以放宽这个要求）。对于给定的 K，算法首先给出一个初始的分组方法，之后通过循环迭代的方法不断对分组进行调整，使每一次迭代后的分组结果都较前一次分组同组中的记录相似度越来越大，不同组中纪录的相似度越来越小。使用划分法基本思想聚类的算法主要有：K-Means 算法、K-Medoids 算法、Clarans 法，其中 K-Means 算法是最常用的一种聚类分析算法[15]；

②层次法(hierarchical methods)：这对给定的数据集进行层次似的分解，直到满足某种条件为止。具体又可分为“自底向上”和“自顶向下”两种方案。例如在“自底向上”方案中，初始时每一个数据纪录都组成一个单独的组，在接下来的迭代中，它把那些相互邻近的组合并成一个组，直到所有的记录组成一个分组或者某个条件满足为止。使用层次法基本思想聚类的算法主要有：BIRCH 算法、CURE 算法等；

③基于密度的方法(density-based methods)：这种方法与其它方法相比有一个根本区别：它不是基于记录之间各种各样的距离（相似度）的，而是基于密度的。这样就能克服基于距离的算法只能发现“圆形”类的缺点。这个方法的指导思想就是，只要一个区域中的点的密度大过某个阀值，就把它加到与之相近的聚类中去。使用基于密度方法基本思想聚类的算法主要有：DBSCAN 算法、OPTICS 算法、DENCLUE 算法等；

④基于网格的方法(grid-based methods)：对给定的数据空间首先将其划分成为有限个单元（cell）的网格结构，所有的处理都是以单个单元为对象。这种方法的一个突出的优点就是处理速度很快，通常的处理速度只与把数据空间分为多少个单元有关，而与给定数据空间中记录的个数无关。使用基于基于网格方法基本思想聚类的算法主要有：STING 算法、CLIQUE 算法、Wave-Cluster 算法；

⑤基于模型的方法(model-based methods)：对给定的数据集为其中每一个聚类假定一个模型，然后在数据集中寻找能个很好的满足这个模型的数据。可以基于数据点在空间中的密度分布函数或其他原则构建这样的数据模型。基于模型的方法有一个潜在的假定：目标数据集是由一系列的概率分布所决定的。通常有两种尝试方向：统计的方案和神经网络的方案。