二阶段招聘信息检索方法*

王静帆[1]，夏云庆[2]，郑方[2]，邬晓钧[3]
(1. 清华大学计算机系 北京 100084; 2. 清华大学信息技术研究院语音和语言技术中心 北京 100084;
3. 清华信息科学与技术国家实验室 北京 100084)

文　摘：招聘信息检索与传统信息检索存在较大差异，传统检索方法不能实现良好的招聘信息检索效果。为解决该问题，本文提出二阶段招聘信息检索方法，针对招聘信息的标题文本和职位描述文本分两阶段分别进行不同的处理。第一阶段本文采用 VSM 模型对标题文本进行初步检索，将相关度较高的招聘信息视为种子；第二阶段，本文采用文本相似度度量方法和聚类分析方法，在招聘信息全集中寻找与种子相似度较高的招聘信息。通过结合“请求-文档”相关度和“文档-文档”相似度，最终计算相似招聘信息与搜索请求的相关度，完成检索结果综合排序。实验证明，这个方法能有效提高招聘信息检索系统的性能。

关键词：招聘信息检索；文本相似度；特征选择
中图分类号： TP391

1　引言

搜索引擎现已成为互联网信息操作的关键入口。典型的搜索流程是这样的：首先用户在搜索引擎用户界界面上输入表达搜索意图的搜索请求（通常是几个关键词）；接下来搜索引擎根据搜索请求包含的关键词进行文档检索；最后搜索引擎将相关文档的链接根据某种相关度依次排列在搜索结果页面上，供用户详细访问。

招聘信息检索具有同普通信息检索相同的目的，即将相关招聘信息页面链接提供给搜索用户。但实际上，招聘信息检索具有自己的独特性。第一，招聘信息检索通常以职位名称作为检索关键词，然而多数职位名称在语言中有多种表达。例如，“经理”职位跟“主管”、“总监”、“主任”等职位名称具有非常接近的职位定义。研究中我们发现，“美工”职位同九种不同的职位名称具有接近的定义。这一情况导致基于关键词的信息检索方法严重失效。显然，检索“经理”职位时，“主管”职位信息不可能获得。第二，招聘信息具有基本的文本结构，包含至少“标题（title）”和“职位描述（description）”两个域。这同普遍意义下信息检索所面向的非结构化文档有所不同。另外一个不同更为明显，即两个域中的文本从实词词汇角度看，几乎不存在重叠。也就是说，出现在标题域文本中的职位名称，几乎不会在职位描述域中出现。这就给基于向量空间模型（VSM[1]）的传统信息检索方法带来麻烦。例如，搜索“经理”这一职位，职位描述域中是不存在这个词汇的，使用 VSM 模型计算搜索请求同文档的相似度，检索效果比较差。我们会在实验部分验证这一事实。

那么是不是说，职位描述域的文本对招聘信息检索没有意义了呢？当然不是。从招聘信息的特点看，仅仅依赖 VSM 模型无法获得较高的检索性能。

本文提出二阶段的招聘信息检索方法，其基本思想是：首先采用 VSM 模型进行初步检索，然后利用文档相似度将相关度靠前的相关职位信息作为种子进行近似扩展。也就是说，利用 VSM 获取相关度较大的文档，在招聘信息库中寻找类似的招聘信息，从而实现检索结果的充分扩展。我们在第一阶段采取 VSM 模型进行初步检索，因此我们的问题集中在通过计算文档相似度获取类似招聘信息上。本文实现了基于 VSM 模型的文档相似度计算方法和基于聚类的相似度改进方法。实验结果显示，第二阶段的工作在不同情况下提升了 0.01 到 0.08 的检索准确率。

本文组织如下：第二节中我们简要介绍招聘信息检索系统的应用背景，通过分析目标文本定位关键研究难题；第三节描述二阶段招聘信息检索方法，重点介绍利用文档相似度和文本聚类对初步检索结果的扩展；我们在第四节对这一方法进行了详尽评测和讨论；第五节我们介绍了相关工作；第六

*基金项目：清华大学基础研究基金(No. JC2007049)
作者简介：王静帆（1982），女（汉），福建，在读硕士生。
通讯联系人：夏云庆，助理研究员，yqxia@tsinghua.edu.cn

节对文章进行总结。

2 招聘信息检索

本文研究的最终目标是开发高性能的招聘信息检索系统。该系统利用爬行器从互联网获取即时的招聘信息，通过一个信息检索界面为用户提供招聘信息检索服务。以下是两条招聘信息实例。

职位信息 1:

<title>软件工程师</title>

<description>熟练掌握 Java, j2EE；熟悉 Eclipse 插件开发优先；富有激情；良好的团队合作精神,中英文流利</description>

职位信息 2:

<title>程序员</title>

<description>须有 Java 开发经验；有 Eclipse 插件开发，RFT 使用经验者优先；计算机或相关专业研究生</description>

这两个实例的职位名称只出现在 title 域中。而用户通常喜欢直接利用职位名称作为关键词进行搜索。这样，对于传统的信息检索方法来说，有时候只有标题域文本是有价值的，而无法利用职位描述域文本。

另一方面，虽然上面两个招聘信息在实际意义层面上非常近似，但是由于他们在名称上无法匹配，利用 VSM 模型无法同时找到两条职位信息。在招聘信息检索中，设法得到近似或相关的职位，提高召回率，对招聘信息检索用户具有重要价值，也是研究面临的重要大挑战。

实际上，以上两个实例也给我们提供了重要线索，即两个职位描述域的文本具有非常高的文本相似度。我们可否做这样的假设：职位描述域文本相似度高的两个职位在意义上是近似的。我们的观察和实验证明了这一假设。

我们的招聘信息检索系统首先利用传统 VSM 模型对检索请求进行检索，获得初步检索结果；接下来，系统对相关度较大的相关招聘信息，利用文档相似度和文本聚类对初步检索结果进行扩展。

3 二阶段招聘信息检索方法

3.1 二阶段思想与混合检索模型

二阶段招聘信息检索方法的基本思想如下：

（1）我们将招聘信息看作由两个域组成的结构化文本，并对这两部分文本进行不同的处理。

（2）我们将“搜索请求-文档”相关度与“文档-文档”相似度结合起来，以发现更多职位名称不同意义接近的招聘信息。

我们的招聘信息检索系统的目标通过两步实现：第一步，采用传统 VSM 模型获取初步的招聘信息。第二步，选择初步招聘信息中相关度最靠前的招聘信息作为种子，通过计算与种子文档的文档相似度获取对近似文档。最终相似文档同搜索请求的相关度计算如下：

$$rel^*(q,d) = \max_{d_i}\{rel(q,d), rel(q,d_i)\times sim(d,d_i)\} \quad (1)$$

这里，$rel(q,d)$ 表示文档 d 与搜索请求 q 的相关度（第一阶段计算得到），$sim(d,d_i)$ 表示文档 d 和文档 d_i 的相似度（第二阶段计算得到）。

3.2 基于 VSM 模型的第一阶段招聘信息检索

在第一阶段，我们采用传统 VSM 模型对招聘信息进行初步检索，以获得搜索请求与文档的相关度。将相关招聘信息按照相关度进行排序，选中位置靠前的相关文档进行第二阶段的扩展。

计算招聘信息同搜索请求的相关度时，有两种思路处理招聘信息文本，即仅适用标题文本和使用招聘信息全文，直觉上后者能取得更好的效果。但是从前面的分析看，有的招聘信息只在标题包含了职位名称，而职位描述文本几乎没有提到这个名称。这种情况下，文档文本量不但不能提高效果，反而可能引入错误。我们将在第四节的实验部分对这两种思路进行评测。

第一阶段检索我们采用了两种“搜索请求-文档”相关度计算方法，即余弦（cosine）和内积（inner product）[2]。在特征选择上，我们采用最常用的词汇特征。词汇特征的权重采取 *tf-idf* 方法计算如下：

$$w_{ij} = tf(t_i, d_j) * \log(N / n_i) \quad (2)$$

这里，w_{ij} 表示特征 t_i 的权重，$tf(t_i,d_j)$ 表示特征 t_i 在文档 d_j 中的出现次数，N 表示所有文档的个数，n_i 表示包含特征 t_i 的文档个数。

最后，通过相关度计算，我们获得与搜索请求相关的招聘信息。在第二阶段，我们将相关度靠前的招聘信息作为种子，通过文档相似度计算获取更多的近似招聘信息。

3.3 相关种子的扩展方法

在第二阶段，我们采用招聘信息全文构建 *tf-idf* 特征向量，使用多种文档相似度计算方法计算相关种子同其它招聘信息的相似度。这里，我们的方法主要涉及如下研究内容：特征选择和相似度计算。

3.3.1 基本的特征和相似度度量

我们尝试采用两类特征，即词汇和汉字二元文法（bi-gram）。 Li 等[2]曾经对用于文本分类的特征进行了评测，结论是词汇和汉字二元文法效果最

好。这是我们采用这两类特征的实践依据。

我们从招聘信息文本集中获取文本特征。在词汇特征处理中，我们首先对停用词进行过滤，最终我们得到了大约 25,000 个词汇特征。汉字二元文法特征具有更高的维度，我们获得了 140,000 个汉字二元文法特征。

我们采用了两种文档相似度度量方法，即余弦（cosine）相似度和扩展 Jaccard 相似度。这两个方法在 Strehl 和 Mooney [6]的评测中取得最好效果，并在中文文本相似度度量中广泛应用。假定两个文档 d_1 和 d_2 对应两个特征向量 v_1 和 v_2，文档 d_1 和 d_2 的余弦距离公式如下：

$$\cos(d_1, d_2) = \frac{d_1 \bullet d_2}{|d_1| \times |d_2|} \tag{5}$$

扩展 Jaccard 相似度公式如下：

$$Jaccard(d_1, d_2) = \frac{d_1 \bullet d_2}{|d_1|^2 + |d_2|^2 - d_1 \bullet d_2} \tag{6}$$

需要指出的是，还有其它相似度度量方法，例如内积和欧氏距离。这些方法对向量计算得到相似度绝对值，不容易进行规一化处理，因此在进行相似文档选择时，很难设定一个可用的阈值。所以，

我们的系统不采用这些方法。

3.3.3 特征选择和降维

VSM 模型的显著特征是高维度，将导致向量空间的高度稀疏问题。在文本分类中，数据稀疏问题通过特征降维方法实现。Yang [4]证实，特征选择能过滤掉大量不携带信息的特征，因此能提高文本分类的性能。在我们的工作中，采用了两种特征选择方法：*DF*（文档频率）和 χ^2 统计量。

DF 表示包含该特征的文档个数。我们保留 DF 评分超过事先设定阈值的特征。

χ^2 统计量最初用于评估特征同分类信息的独立程度。在特征选择中，独立程度低的特征将被过滤掉。由于 χ^2 统计量是针对监督学习环境下的特征选择方法，因此需要产生文档类别。我们通过聚类技术达到这一目的。这里采用了重复二分聚类（repeated-bisection clustering）[5]方法自动建立文本类别。基于 χ^2 统计量计算如下：

$$\chi^2(t, c) = \frac{N(AD - BC)^2}{(A+C)(B+D)(A+B)(C+D)} \tag{3}$$

其中，*A* 表示类别 *c* 中包含特征 *t* 的文档个数，*B* 表示类别 *c* 中不包含特征 *t* 的文档个数，*C* 表示不在类别 *c* 中但包含特征 *t* 的文档个数，*D* 表示不在类别 *c* 中也不包含特征 *t* 的文档个数，*N* 表示所有文档的个数。每个特征的评分定义为：

$$\chi^2_{\max}(t) = \max_c \{\chi^2(t, c)\} \tag{4}$$

3.3.4 聚类分析和相似度重估

以上，两个文本之间的相似度仅考虑他们之间出现的公共特征。这种相似度和人们关注的“相似”还有一定的偏差。首先，有多种行文方式来表达相同语义，对于招聘信息这样的短文本（大部分在 100 字以下），稀疏问题仍然比较严重。同时招聘文本中有的信息也非一般用户关心的（一般用户不会关注“两年以上工作经验”，“待人诚恳”，用户往往更关注职业信息的专业，领域等），这些共同词汇可能造成两个领域无关的职业描述相似度过高。在这里，我们引入聚类分析方法来修正文本相似度。聚类算法综合考虑了多个文本之间的相似度，可以修正之前单纯考虑两个文本同现词语可能引入的错误。聚类算法的目的是将相似的文本聚集在一个簇中，因此，我们认为在相同的簇中的文本具有较高的文本相似度。

具体的方法是：首先通过聚类方法将文本划分成固定的 *k* 个簇。在计算两个文本的相似度时，先采取基本方法的计算出相似度，然后根据聚类结果对该相似度进行重估。如果在严格聚类结果中某文档同种子文档同属一个类，则将文档相似度进行一定程度的提升。公式如下：

$$\begin{aligned} &e\_sim(d_1, d_2) \\ &= \alpha \times b\_sim(d_1, d_2) + \beta \times is\_in\_cluster(d_1, d_2) \end{aligned} \tag{7}$$

其中，$b\_sim(d_1, d_2)$ 表示基本相似度，$e\_sim(d_1, d_2)$ 表示扩展相似度，$is\_in\_cluster(d_1, d_2)$ 表示文档 d_1 和 d_2 是否属于同一类，α 和 β 是权重常量。在实际计算中，我们采取 0.8 和 0.2。我们依然采用 CLUTO[5]中的重复二分聚类方法，将左右招聘信息文档天然的划分为 *k* 个簇。

4 实验与评测

4.1 实验设置

数据

我们的招聘信息库包含大约 55,000 条中文招聘信息，来自中华英才网（http://www.chinahr.com）和无忧工作网(http://www.51jobcom)。每条招聘信息包含标题和职位描述。我们的测试集包含 100 个搜索请求。

评测指标

我们采用了特定数目的搜索结果的准确率作为评测指标，即 p@N，在获取到前 N 个中的结果中相关文档所占比例。如 p@10=0.4 表示，对查询获得的前 10 个结果，有 4 个被认为是相关的。其

中，对 N 取了 1、5、10、20、30 和 40 作为实际的限定数目。我们对 55,000 条招聘信息中的 5000 条进行了手工标注,作为测试集。实验结果为在这 5000 测试集合上的 p@N。

4.2 实验 1：第一阶段检索

这个实验的目的是评测第一阶段招聘信息检索的效果。在检索目标上，我们面向两种文本：标题域文本和招聘信息全文。这样的设置是为了考察招聘信息职位描述域文本对第一阶段检索的贡献。

由于采用了 VSM 模型，需要进行特征选择。我们采用词汇特征，采用 *tf-idf* 作为特征权重。在计算招聘信息与搜索请求的相关度时，我们采用了余弦距离和内积方法。实验结果如图 1 所示。

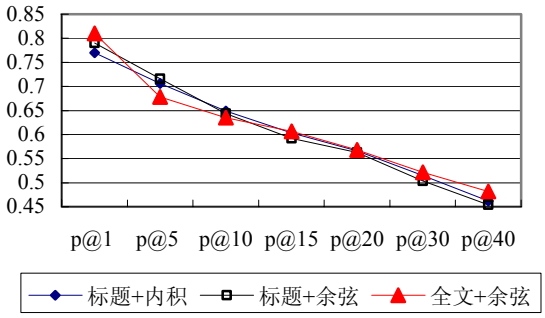

图 1　第一阶段 *p@N* 曲线：比较标题域和招聘信息全文在进行余弦距离和内积相关度计算上的性能差异。p@*N* 表示前 *N* 个搜索结果的准确率。

实验结果可得到两个结论：第一，采用招聘信息全文进行招聘信息检索的效果比仅采用招聘信息标题域文本几乎没有什么提高。经过分析我们发现原因主要是搜索请求基本上是职位名称，而有的职位名称只在招聘信息标题文本中出现，不在职位描述文本中出现，这是导致采用全文不能提高效果的根本原因。这样的实验结果启发我们考虑对职位描述文本采取另外的利用方式。在后面的实验中，我们将第一阶段获得的相关招聘信息中相关度最高的几条招聘信息看作种子，通过计算其它招聘信息在全文上的相似度获得近似招聘信息。

第二，在“请求-文档”相关度计算中，面向标题文本的内积方法性能稳定。我们将这个结果视为后续实验的基线（baseline）方法。

4.3 实验 2：文档相似度与相关招聘信息扩展

本实验中我们分三步评测文档相似度对相关招聘信息的扩展效果。我们首先评测不同的特征选择对文档相似度计算的影响。我们实现了词汇特征和汉字二元文法两类特征。由于特征选择方法和相似度计算方法在后面评测，这里我们暂时不进行特征选择，并采用余弦距离进行相似度计算。实验结果如图 2 所示。

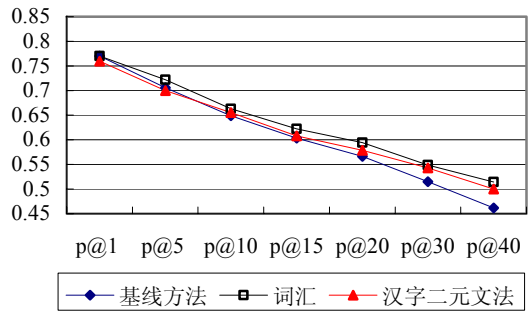

图 2　通过文档相似度计算对基线方法改进的 p@N 曲线：比较词汇和汉字二元文法两类特征。

从图 2 中的实验结果可以得到两个重要结论：第一，在我们的评测中，通过文档相似度对相关招聘信息进行扩展，准确率都有一定程度的提高。同时从 p@1 到 p@40 的提高程度不断增大，在 p@40 上的提高达到了 0.052。这表明，测试范围越大，扩展方法的效果越明显。方法在 p@1 上几乎没有提高，具体到某些单个请求，效果反而下降。这是因为，我们的扩展是建立在第一阶段检索得到的种子文档基础之上。扩展效果的严重依赖于种子文档与请求的相关度。例如，某些请求在 p@1 中产生种子文档实际上是错误的，这必然导致扩展的错误。

第二，词汇特征效果在所有情况下都略好于汉字二元文法。这同[3]在文本分类中的评测结果略有差异，他们的结论是汉字二元文法略好。这可能同招聘信息文本的词汇特点有关，招聘信息中的特征维度较小，同时超过两个汉字的词汇特征数量较多。

接下来的实验我们对比不同相似度度量方法对检索准确率的影响。我们实现了余弦距离和扩展 Jaccard 方法，采用词汇特征，实验结果如图 3 所示。

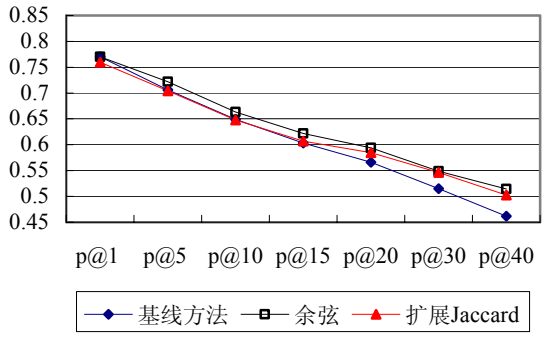

图 3　通过文档相似度计算对基线方法改进的 p@*N* 曲线：比较特征选择方法。

图 3 显示，余弦距离比改进 Jaccard 方法效果略好，准确度相差 0.01~0.02。

4.4 实验 3：特征选择和降维处理

接下来的实验中，我们采用两种方法对现有的

特征进行特征选择，即 *DF* 和 χ^2（Chi）。需要特别指出的是，我们在特征选择时，选择比例没有固定的数值，对特征和特征选择方法的每个组合，我们只展示取得相关文档扩展效果最好的组合。例如，对于词汇与 *DF* 的组合，在保留 20%的特征时，相关文档扩展达到最好。实验结果如图 4 所示。

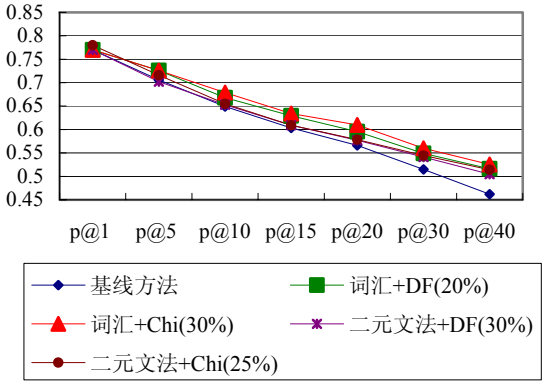

图 4　通过文档相似度计算实现对基线方法的改进 p@*N* 曲线：比较特征选择方法。

图 4 的实验结果显示，在词汇特征上采用 χ^2 特征选择方法保留 30%的词汇特征时，相关文档的扩展效果最好。同采用全部特征相比，检索准确率并没有下降。

4.5　实验 4：聚类与相关招聘信息扩展

本实验我们测试聚类分析对相似度进行重估后的效果。我们以实验 2 中最好结果为基础，即“词汇+余弦距离”方法。采用 CLUSTO[5]获取聚类结果，对上述方法的结果进行重估，实验结果如图 5 所示。

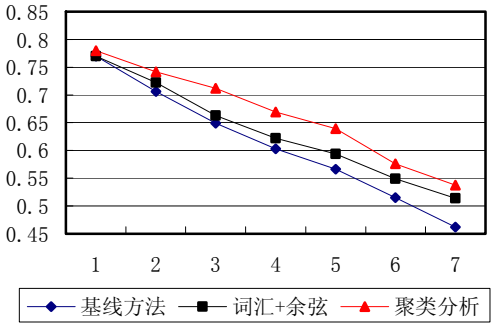

图 5 采用聚类分析方法修正文档相似度后的 p@*N* 曲线。

如图 5 显示，聚类方法的引入，对基线方法在不同情况下有 0.04~0.07 的提高，对“词汇+余弦距离”方法有 0.01 到 0.05 的提高。从而证明了这个方法是有效的。

4.6　实验 5：招聘信息检索系统

最后，我们将上述实验中取得较好效果的特征、特征选择方法、相似度计算方法和聚类方法融合到我们的招聘信息检索系统中。即，对招聘信息文本，我们采用词汇作为特征构建 *tf-idf* 向量，通过 Chi 方法进行特征选择，利用余弦距离进行文本相似度计算，然后再用聚类结果对文本相似度进行重估；最后我们将第一阶段的相关度和第二阶段的相似度融合，获得相似招聘信息与搜索请求的相关度。实验结果如图 6 所示。

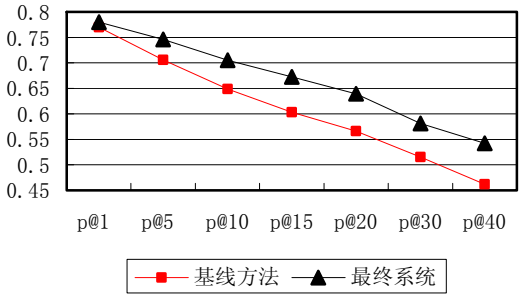

图 6　招聘信息检索系统对基线系统的提高。

可以看出，最终系统相对于基线方法在不同情况下得到了 0.01 到 0.08 的提高。

5　相关工作

本文提出的二阶段招聘信息检索方法受到查询请求扩展技术[7][8]的启发。这些方法或者从最靠前的相关文档中自动抽取有用词汇，或者从词典中获得查询请求词汇的近义词，然后利用这些词汇构建一个扩展的搜索请求，利用扩展的搜索请求进行检索，将检索结果提供给检索用户。查询请求扩展在某些特殊应用情况下非常必要，即第一次查询能反馈的相关文档数量很少。招聘信息检索也面临类似的困难，某些职位的招聘信息非常少。查询请求扩展技术对招聘信息检索意义不大，原因是招聘信息的职位描述文本在词汇使用上几乎不与标题文本重叠，利用查询请求扩展技术只能获得价值不大的词汇，例如“工作经历”、“学历”等多数招聘信息都包含的词汇。

基于 VSM 模型的请求-文档相关度度量方法在许多信息检索方法[1][9]中涉及到。不同的是，我们还计算文档-文档相似度。

基于 VSM 模型的文档相似度度量方法在一些文本分类和聚类方法中得到了广泛应用。Li 等[2]在中文文本分类研究中比较了两类特征，即词汇和汉字二元文法。Yang 和 Pedersen[4]评测了五类特征选择方法，即 DF, IG (信息增益), CHI(χ^2 统计量), TS (术语强度) and MI (互信息)。其中 DF 和 CHI 方法效果最好。在我们的工作中采用了这些方法。利用聚类结果作为类别标签计算 CHI 的方法则是来自于 Liu Tao[10]，他们利用 K-means 聚类计算 CHI，用于提高聚类准确度。

6 结论与未来工作

针对招聘信息检索与传统信息检索存在较大差异所导致的传统检索方法不能良好完成招聘信息检索的问题，本文提出二阶段招聘信息检索方法，对招聘信息的标题文本和职位描述文本分两阶段分别进行不同的处理。第一阶段本文采用 VSM 模型对标题文本进行初步检索，将相关度较高的招聘信息视为种子；第二阶段，本文采用文本相似度度量方法和聚类分析方法，在招聘信息全集中寻找与种子相似度较高的招聘信息。通过结合“搜索请求-文档”相关度和“文档-文档”相似度，最终计算相似招聘信息与搜索请求的相关度，完成检索结果综合排序。实验证明，这个方法能有效提高招聘信息检索系统的性能。

在实际系统中，由于两个文本之间的相似度跟用户输入查询无关，因此我们可以离线计算这种相似度，对每个文本保留 N 个近邻。在用户查询时，可以直接扩展出相关近邻和相似度评分，以节省查询时间。我们的系统的特点在于，利用了招聘信息文本中存在的结构关系和文本特点。在未来的研究中，我们将进一步细化文本中可能存在的结构，利用信息抽取技术获取结构信息，进一步提高招聘信息检索性能。

参 考 文 献

[1] R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. 1999. Addison Wesley/Pearson.

[2] Salton and Buckley Term-weighting approaches in automatic text retrieval. Information Processing & Management 24,5,513~523, 1988.

[3] Jingyang Li, Maosong Sun and Xian Zhang, A comparison and semi-quantitative analysis of words and character-bigrams features in Chinese text Categorization. ACL06.

[4] Yang, Yiming and J. O. Pedersen. A comparative study on feature selection in text categorization. In Proceedings of the 14th International Conference on Machine Learning (ICML-97),1997.412～420.

[5] Karypis G. CLUTO. A Clustering Toolkit. Dept. of Computer Science, University of Minnesota, May, 2002.

[6] A Strehl, J Ghosh and R Mooney. Impact of similarity measures on web-page clustering. In Proc. AAAI Workshop on AI for Web Search, Austin, 2000io.

[7] C. Buckley, G. Salton, J Allen and A. Singhal. Automatic query expansion using SMART: TREC 3. In the 3rd Text Retrieval Conference,69~80,1994.

[8] C Carpineto, R de Mori, G Romano, B Bigian. Information Theoretic Approach to Automatic Query Expansion. ACM trans. on Information System.2001: 19(1):1～27.

[9] Justin Zobel and Alistair Moffat. Exploring the similarity space. SIGIR’1998

[10] 刘涛 ，吴功宜 ，陈正. 一种高效的用于文本聚类的无监督特征选择算法. 计算机研究与发展 2005, 42(3),381~386
Liu Tao , Wu Gongyi , Chen Zheng. An effective unsupervised feature selection method for text clustering. Journal of computer research and development. 2005,42(3),381~386 [in Chinese]

A Two-step Job Information Retrieval Method

Jingfan Wang[1]，Yunqing Xia[2]，Fang Zheng[2]，Xiaojun Wu[3]

(1. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China;
2. Center for Speech and Language Technologies, Research Institute of Information Technology, Tsinghua University, Beijing 100084, China;
3. National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract：Job information retrieval holds unique characteristics compared to traditional information retrieval leading to lower performance. In this paper, text structure of the job posting is analyzed and a two-step framework is proposed to address this problem. In the first step, the traditional vector space model (VSM) model is applied on the title field to get relevant job postings. In the second step, the top relevant documents are considered as seeds to retrieve similar job posting using document similarity measures and clustering technique on full text. Finally, relevant scores between the similar job postings and the query are calculated by combing relevant scores in the first step and similarity scores in the second step. Our experiment proves that this method is effective in job information retrieval.

Key words：Job information retrieval; document similarity; feature selection