中文分词在网络招聘个性化职位推荐算法的应用

摘要：本文探讨了中文分词在网络招聘个性化职位推荐中的应用。通过介绍基于词典、统计和理解的分词方法，以及mmseg消除歧义的规则和mmseg4j工具，本文详细分析了中文分词技术在处理招聘信息中的关键作用。接着，本文阐述了中文分词在提取求职者与职位信息、计算相似度以及推荐匹配职位等方面的具体应用。本文认为，中文分词技术能有效提升网络招聘的个性化推荐效果，为招聘双方提供更高效、精准的匹配服务。

关键字：中文分词；mmseg；网络招聘；算法推荐

Abstract：This article explores the application of Chinese word segmentation in personalized job recommendation for online recruitment. By introducing word segmentation methods based on dictionaries, statistics, and comprehension, as well as the mmseg rules for disambiguation and the mmseg4j tool, this article analyzes in detail the crucial role of Chinese word segmentation technology in processing recruitment information. Subsequently, the article elaborates on the specific applications of Chinese word segmentation in extracting job seeker and job information, calculating similarity, and recommending matching jobs. This article argues that Chinese word segmentation technology can effectively enhance the personalized recommendation effect of online recruitment, providing both recruiters and job seekers with more efficient and precise matching services.

Keywords: Chinese Word Segmentation; mmseg; Online Recruitment; Algorithm-based Recommendation

**1.引言**

随着信息技术的迅猛发展和互联网的普及，网络招聘已成为求职者与招聘方的重要桥梁。然而，面对海量的职位信息和求职者信息，如何高效、准确地匹配双方需求，提高招聘效率，成为了亟待解决的问题。中文分词技术作为自然语言处理的基础技术之一，在文本信息处理和挖掘中发挥着重要作用。本文将探讨中文分词技术在网络招聘个性化职位推荐算法中的应用，以期实现更加精准、高效的职位推荐，满足求职者和招聘方的需求。

**2.中文分词相关研究**

在大模型中，分词算法的选择和优化非常重要。通常会选择一些高效的、准确的分词算法，并结合实际应用场景进行优化和调整，以确保模型的性能和效率。分词是大模型中的关键预处理步骤之一，它直接影响着模型的性能和效率。分词算法也需要与模型的训练和推理过程紧密结合，以确保模型的准确和稳定。分词算法在大模型中具有重要的地位，它是自然语言处理中的基本任务之一，直接影响着大模型的性能和效率。中文文本之间没有界限，所以中文的语义理解的第一步是进行专门的分词处理。由于基于字符串的分词法简单高效，应用较广，所以我们本文使用字符串匹配分词法。

目前的中文分词器，基本都是这三类的任意组合，一般来说，只使用其中一种的分词器效果基本无法达到比较理想的分词效果。而在这些组合中，又以前两种的组合居多，这可能也与第三种方法本身的研究进展和难度系数相对较高有关。

**2.1 三种分词方法**

**2.1.1基于词典与规则的分词方法**

基于词典与规则的分词方法主要是利用一个预先定义好的词典（也称为“词表”或“词库”），将待处理的文本中的字符串与词典中的词条进行匹配。如果文本中的某个字符串与词典中的某个词条一致，则将其作为一个词切分出来。

主要方法为：（1）正向最大匹配法（MM法）：假定分词词典里的最长词的长度为k，那么就用待处理文档的当前长度为k的字符串和词典进行匹配。如果找到则将分词，否则去掉最后一个字符再进行匹配，直到成功分词或剩余字符串长度为0。例如，词典中的最长词长度为4，待处理文本为“研究生命的起源”，则首先尝试匹配“研究生”，若词典中存在该词，则成功分词；若不存在，则尝试匹配“研究生命”，依此类推。

（2）逆向最大匹配法（RMM法）：与正向最大匹配法类似，但匹配方向相反，从待处理文本的末尾开始向前匹配。这种方法在某些情况下能提高分词的准确性，尤其是在处理歧义词时。

（3）双向最大匹配法（Bi-direction MM法）：同时使用正向最大匹配法和逆向最大匹配法进行分词，然后比较两种方法的分词结果。如果两种方法的分词结果一致，则认为分词正确；否则，通常按照某种策略（如选择分词数量较少的结果）进行决策。

1. 其他规则：除了上述的基本匹配方法外，还可以结合其他规则来提高分词的准确性。例如，可以定义一些特殊符号（如标点符号、空格等）作为分词标记；也可以定义一些固定的词组或短语作为不可分割的整体。

根据李晓东的统计表明，单纯使用正向最大匹配的错误率为1／169，单纯使用逆向最大匹配的错误率为1／245。（李晓东，2008：9）

**2.1.2基于统计的分词方法**

基于统计的分词方法认为，如果相邻的字在文本中出现的次数越多，那么它们就越有可能构成一个词。因此，字与字相邻出现的概率或频率可以较好地反映成词的可信度。方法步骤为：

1. 收集并预处理语料库：语料库是包含大量文本数据的数据库，用于训练和测试分词模型。要对语料库进行收集、清洗和预处理，以便后续的分析和建模。

（2）构建统计模型：建立一个概率模型，用于描述每个中文字在不同上下文环境中出现的概率。

典型的模型包括隐马尔可夫模型（HMM）、最大熵模型（ME）、条件随机场（CRF）等。

（3）分词：在实际应用中，分词通常采用正向最大匹配法（FMM）和逆向最大匹配法（RMM）等方法。根据前面构建的统计模型和特定的分词策略，将待处理的文本分割成词语序列。

（4）评估和优化：分词完成后，需要对结果进行评估和优化。通常采用F1值、准确率、召回率等指标来评价分词的效果，并对模型进行调整和改进。

在基于统计的分词方法中，条件随机场（CRF）和隐马尔可夫模型（HMM）是两种常用的模型。CRF算法可以更加准确地建立标签之间的依赖关系，而HMM算法则通常基于字或词的频率统计建立模型。这两种模型在中文分词领域都有广泛的应用。

**2.1.3基于理解的分词方法**

**其**基本思想是在分词的同时进行句法、语义分析，利用句法信息和语义信息来处理歧义现象。它模拟了人脑在分词过程中对句子进行理解的过程。方法步骤为：

（1）分词子系统：负责将待处理的文本进行初步的词语切分。

（2）句法语义子系统：对分词子系统切分出的词语进行句法、语义分析，获取关于词、句子等的句法和语义信息。

（3）总控部分：在总控部分的协调下，分词子系统可以获得句法语义子系统提供的信息，从而对分词歧义进行判断。

基于理解的分词方法主要有两种技术路线：

（1）专家系统分词法：

从专家系统角度将分词的知识（包括常识性分词知识与消除歧义切分的启发性知识即歧义切分规则）从实现分词过程的推理机中独立出来。使知识库的维护与推理机的实现互不干扰，从而使知识库易于维护和管理。具有发现交集歧义字段和多义组合歧义字段的能力和一定的自学习功能。

（2）神经网络分词法：

利用神经网络模型来模拟人脑对句子的理解过程，从而进行分词。通常需要大量的训练数据来训练模型，以达到较好的分词效果。

基于理解的分词方法虽然在目前还处在试验阶段，但其潜力巨大。随着人工智能和自然语言处理技术的不断发展，该方法有望在未来得到更广泛的应用。它特别适用于那些对分词准确性要求较高的场景，如机器翻译、情感分析、智能问答等。在这些场景中，准确的分词结果对于后续任务的效果至关重要。

**2.2 MMAEG**

**2.2.1mmseg消除歧义的四大规则**

以很常见的mmseg机械分词算法为例，mmseg在搜索引擎Solr中经常使用到，是一种非常可靠高效的分词算法。mmseg消除歧义的规则有四个，它在使用中依次用这四个规则进行过滤，直到只有一种结果或者第四个规则使用完毕。这个四个规则分别是：

（1）最大匹配，选择“词组长度最大的”那个词组，然后选择这个词组的第一个词，作为切分出的第一个词，如对于“中国人民万岁”，匹配结果分别为：

中/国/人

中国/人/民

中国/人民/万岁

中国人/民/万岁

在这个例子“词组长度最长的”词组为后两个，因此选择了“中国人/民/万岁”中的“中国人”，或者“中国/人民/万岁”中的“中国”。

（2）最大平均词语长度。经过规则1过滤后，如果剩余的词组超过1个，那就选择平均词语长度最大的那个(平均词长=词组总字数/词语数量)。比如“生活水平”，可能得到如下词组：

生/活水/平 (4/3=1.33)

生活/水/平 (4/3=1.33)

生活/水平 (4/2=2)

根据此规则，就可以确定选择“生活/水平”这个词组

（3）词语长度的最小变化率，这个变化率一般可以由标准差来决定。比如对于“中国人民万岁”这个短语，可以计算：

中国/人民/万岁(标准差=sqrt(((2-2)^2+(2-2)^2+(2-2^2))/3)=0)

中国人/民/万岁(标准差=sqrt(((2-3)^2+(2-1)^2+(2-2)^2)/3)=0.8165)

于是选择“中国/人民/万岁”这个词组。

（4）计算词组中的所有单字词词频的自然对数，然后将得到的值相加，取总和最大的词组。比如：

设施/和服/务

设施/和/服务

这两个词组中分别有“务”和“和”这两个单字词，假设“务”作为单字词时候的频率是5，“和”作为单字词时候的频率是10，对5和10取自然对数，然后取最大值者，所以取“和”字所在的词组，即“设施/和/服务”。

**2.2.2 mmseg4j**

mmseg4j是使用mmseg算法实现的中文分词器，mmseg4j分词器常见的分词方法有三种，分别是simple、complex和max-word。

（1）simple根据最大前向匹配原则(FMM)执行分词匹配[36]。最大前向匹配算法是按从左到右顺序从词典中找到最长的匹配词前缀，如果找到最长匹配单词，则把 这个词从要删除的分词文本中删除；如果没有，则删除要分割文本的最后一个单词以形成新的分割文本，然后与字典中的词条进行匹配，重复直至分词文本中只 剩最后一个字符，并将其作为单词切分。

（2）complex将切分的相邻三个词作为一个词块，在最大匹配原则基础上增加了四个过虑规则，用以消除歧义，规则如下：

侯选词块的词长之和最大； 侯选词块的平均词长最长； 侯选词块的词长变化最小； 侯选词块的单字出现词自由度最高。

（3）max-word分词方法比前两者都要全面，它会对文本执行最细粒度的分词， 找出所有可能的组合，本文使用max-word分词方法作为系统索引的分词方式mseg4j是使用mmseg算法实现的中文分词器，mmseg4j分词器常见的分 词方法有三种，分别是simple、complex和max-word。

**3.中文分词在网络招聘个性化职位推荐算法的应用**

在进行人岗匹配任务时，文本预处理是一个必不可少的步骤。职位需求与求职者简历经历中的文本数据通常包含大量的噪声、冗余和不规则性，这些问题会 影响人岗匹配模型的性能，影响到后续的处理和分析，且容易造成资源浪费等情 况，因此，在进行模型训练之前，需要先对数据进行预处理操作，通过文本预处理，可以提高文本数据的质量和可用性，从而为后续的人岗匹配任务提供更加准 确和可靠的基础。

**3.1 数据清洗**

原始数据中，可能会包含一些不完整、无意义的词语和符号，比如：“￥”、 “#”、“@”、“&”等非文本信息和连词、助词等常用功能词。这些数据不 仅会使数据维度增加，造成资源浪费，而且可能会使模型效果达不到预期效果，影响后续的分析，因此需要对这类数据进行清洗，清洗方法如下：

（1）去重。将职位要求和简历经历中的重复项删除，仅保留首次出现的记录；

（2）正则表达式替换。将职位描述中的非法字符，如“?!~[]\'"@#$%”等符号、“岗位职责，职位描述，工作内容，岗位描述”等常用功能词替换为空；

（3）无效数据去除。将职位描述中一些职位描述为空、无效的空格或有违常理的数据进行删除。

（4）外文文本翻译。调用翻译软件将职位描述为英文的数据翻译为中文文本。

**3.2清洗后进行文本分词**

**3.2.1简历信息的解析**

（1）基础信息提取：通过中文分词技术，可以从求职者的简历中提取出基本信息，如姓名、教育背景、工作经历等。这些信息是构建求职者画像的重要基础。

（2）深层次信息挖掘：除了基础信息外，中文分词还能帮助解析简历中的项目经验、技能描述等文本内容，从而挖掘出求职者的专业技能、职业偏好、工作风格等深层次信息。

**3.2.2职位信息的解析**

（1）岗位需求识别：企业发布的职位信息通常包含职位描述、技能要求、工作内容等关键信息。通过中文分词技术，可以准确地识别这些需求，为后续的职位匹配提供基础。

（2）企业文化解析：职位描述中可能还隐含着企业的文化和价值观。中文分词结合自然语言处理技术，有助于捕捉这些非显性的信息，使职位推荐更加贴合企业的实际需求。

**3.2.3个性化推荐算法的优化**

（1）精准匹配：基于中文分词技术提取的简历和职位信息，可以构建更加精准的匹配模型。例如，通过分析求职者的技能专长和企业的技能要求，可以实现技能的精确匹配。

（2）推荐效率提升：通过中文分词对大量简历和职位信息进行预处理和索引，可以显著提高推荐算法的搜索和匹配效率。

（3）用户体验改善：个性化的职位推荐能够减少求职者筛选职位的时间成本，提高求职效率。同时，精准的推荐结果也能提升求职者的满意度和体验。

**3.2.4实际应用案例**

在实际招聘平台中，如智联招聘、前程无忧等，都采用了基于中文分词的个性化推荐算法。这些平台通过收集和分析大量的简历和职位信息，利用中文分词技术提取关键信息，并构建智能推荐系统，为求职者和企业提供更加精准的匹配服务。

综上所述，中文分词在网络招聘个性化职位推荐算法中扮演着重要的角色。它不仅有助于解析和处理海量的简历和职位信息，还能为个性化推荐算法提供精准的数据支持，从而提升招聘效率和用户体验。

**参考文献**

[1]隋在娟. 基于分词算法的用户个性化推荐系统设计[J]. 数字通信世界,2020(12):115-116.

[2]王洪伟,段友祥. 面向微博的个性化内容推荐算法研究[J]. 计算机与数字工程,2022,50(1):152-156.

[3]李晓东. 搜索引擎中中文分词与纠错模块的设计与实现[D]. 北京:北京交通大学,2008.

[4]蔡宗文,廖丽霞,危福泉,等.基于分词技术的防震减灾知识推荐服务与应用[J].防灾减灾学报,2023,39(02):71-75.

[5]张兴宇.基于用户兴趣度的微博混合推荐算法研究[J].淮南职业技术学院学报,2023,23(04):146-149.