Brief Article

The Author

October 27, 2013

1 Abstract

本篇文章主要介绍了我们小组在今年的Web Track比赛中关于Ad-hoc任务的试验方法。Ad-hoc 任务的目的是从给定的静态网页集合中检索出与查询主题相关的文档，并对文档按照与查询的相关性进行排序。在本文中我们采用的方法是首先对查询词进行扩展，然后使用扩展后的查询词集进行文本检索，并将返回的文档集进行排序。

2 Introduction

Ad-hoc搜索任务是根据给定的topic从大量的静态网页文件中搜索出与该topic最相关的文档。这个过程与用户进行web搜索的过程非常相似，都是从一个大的文本数据集中查找与检索词相关的文本，继而对文本进行排序的过程，而且，Ad-hoc中给定的topic也与用户搜索具有相同的特点，关键词过于短小而且具有歧义性，因此，仅仅通过将这些关键词与文本集进行匹配的方法，很难找出与检索主题最相关的网页文本，针对这个问题，我们采用先对检索词进行扩展，再计算扩展后检索词集与网页文本之间的相似度，最终根据计算出的相似度对返回结果进行排序的方法，实现Ad-hoc任务中文本检索的功能。今年的web track使用的是clue2013数据集。Web clue 是由cmu提供的xxxx抓取了一年的数据。我们的实验采用的是clueB数据集，它是clue2013的一个子集，约500G。本文剩余部分结构如下，第二章主要是介绍数据的预处理，即索引的建立，第三章是基于查询扩展的搜索模型，第四章是实验结果分析，最后一章是总结和展望。

3 Preprocessing

由于数据集过于庞大，直接对数据集进行操作效率太低，因此我们采用先对数据集建立索引的方式。实验中我们首先使用Indri工具作为索引构造器，配合spammer [?]，对数据进行建索引，然后使用indri内部的检索机制对topic检索结果进行一次筛选，再使用全文搜索引擎lucene对筛选出的网页文本进一步建立索引，其后的所有工作都是在第二次建立的索引上进行的。我们这样做的动机是，在我们实验室内部已经建立了一套完整的基于lucene的检索框架和机制，而TREC官方提供的是一份基于indri的压缩数据，而且indri在处理大数据方面较lucene有性能上的优势，因此，我们采用了将两者结合的方式对数据进行预处理过程。而且，通过解析网页原文本我们发现官方提供的数据虽然经过了一定的处理，但是文本中仍然存在大量噪声数据，特别是在网页的body中存在很多

不相关信息，因此，我们在用lucene对网页文本建立索引之前，我们使用了实验室自己开发的基于网页Block结构的正文抽取工具对文本内容进行解析，事实证明，在对数据进行预处理之后再进行相关检索，确实可以提高检索的效果。

4 Search

在搜索策略中，我们分为三个步骤，第一，对各个topic的检索词进行查询扩展，形成新的查询条条件。第二，使用新的查询词进行检索，返回检索结果。最后，对第二步中返回的检索结果进行排序。

1.查询词扩展 查询扩展是当前检索系统常用的一种提高查询准确率和召回率的技术，主要分为全局分析方法、局部分析方法和基于外部语料库的方法等。全局分析方法是将整个语料库作为扩展词的来源，利用聚类等相关技术从整个预料集中查找扩展词，该方法的一个显著缺点是当语料集中文本数量过大时，效率会明显下降，而局部分析方法针对全局分析的弱点，首先使用相关技术得到一部分与查询词相关的文本，经过分析从这些文本中得到扩展词。 在本实验中，我们采用基于外部预料的方法获取扩展词，这样可以消除查找扩展词的过程中对内部语料库的依赖性。在实验中，我们将扩展次来源分为两个部分，一个是使用了元搜索的方法，利用相关的api调用的方法把google search和bing的搜索接口统一调用，对50个topics分别进行搜索并返回其前50项搜索结果。然后使用了基于Block的正文抽取算法[**?**], 得到各个网页的正文内容，并进行文本切词，消除停用词以后，对剩余的文本词汇计算其tf值，再根据它们的tf值进行排序，从而得到初始的扩展词表；另一个是对wikipedia中相关的页面进行锚文本的计算，通过各个锚文本链接的相互关系，使用pageRank的方法得到权重最高的锚文本节点。 在对文本进行切词时，我们使用了stanford的nlp分词软件，将这些词进行词根化和词性标注，仅保留名词作为扩展词。

2.构建查询语句

我们将扩展词中与原查询的同义词或者本身词作为最大项，然后对其他词的词频进行处理得到权重 $w_i = \frac{TF_i}{TF_{max}}$,然后扩展原查询 $q_{expan} = q_{origin} + \sum_{i=1}^{n} w_i * Expan_i$, 将此扩展后的查询作为新的查询。

3.重排序

利用扩展后的查询放到Indri中进行搜索，主要使用BM25语言模型。

5 Experiments

本次由于我们只提交了一个run，在B数据集上的结果效果很不好。

6 Conclusion Future

通过这次实验，我们验证了查询扩展在辅助查询的方面有所帮助。