

基于概念的网页相似度处理算法研究

郭晨娟,李战怀

(西北工业大学 计算机学院, 陕西 西安 710072)

(nicolegcj@gmail.com)

摘 要:针对海量网页信息,提出适于搜索引擎使用的网页相似度处理算法。算法依据网页抽象形成的概念,在倒排文档基础上建立相似度处理模型。该模型缩小了需要进行相似度计算的网页文档范围,节约大量时间和空间资源,为优化相似度计算奠定了良好基础。

关键词:相似网页;概念抽取;聚类分析;消重

中图分类号: TP393.09 **文献标识码:** A

Concept based algorithm of dealing near-replicas of documents on the Web

GUO Chen-juan, LI Zhan-huai

(Department of Computer Science, Northwestern Polytechnical University, Xi'an Shaanxi 710072, China)

Abstract: To solve near-replicas of documents on the Web obtained by search engine, a similarity dealing algorithm was proposed. Based on concepts extracted from the Web pages and inverted file, the algorithm built a model which shrank the scale of the Web pages processed. The algorithm saved a great deal of temporal and spatial resources and provides a good foundation for near-replicas detection.

Key words: near-replicas documents; concept extraction; cluster analysis; near-replicas detection

0 引言

随着越来越多结构庞大且日益复杂网站的出现,如何有效组织和检索网络信息成为当前网络技术研究的重点之一。搜索引擎是解决网络信息收集和检索查询的有效手段,该技术通过一定方式获得网站网页资料,在此基础上建立数据库并对网页进行索引,最后为用户提供查询功能。

网络中存在大量转载网页,即一篇网页文章内容会以相似或者相同的形式出现在其他网页中。搜索引擎在收集网页过程中,必然会收集大量主题内容相似或者相同的网页^[3]。如果不对这些网页进行处理,不但在网页索引时会浪费大量时间和空间资源,而且为用户提供查询功能时也将显示不必要的重复信息。因此在网页信息收集和组织过程中无可避免地需要进行同主题网页消重工作,即网页相似度处理。

网页相似度处理是聚类分析的一种,用于计算网页文档之间的距离。根据聚类分析的各种方法:划分方法、层次方法、基于密度的方法、基于网格的方法和基于模型的方法等,可以将满足相似度要求的文档划分为一个簇^[6]。对于非海量数据或者非频繁动态变化数据,简单采用上述聚类分析方法可以达到较好的聚类效果。但是对于海量的网页信息,以及搜索引擎持续不断收集网页的特点,简单基于上述聚类分析方法进行网页相似性验证将浪费大量时间和空间资源。

针对搜索引擎网页信息收集过程以及转载网页的特点,本文基于网页主题概念,提出一种高效的网页相似度处理方法。该方法根据网页文档形成的概念,从数据库中获取包含这些概念的文档集合以迅速缩小需要进行相似度处理的文档范围,为快速识别数据库中相似的网页、达到网页文档较好聚类效果奠定基础。

1 系统结构

搜索引擎网页文档的消重处理一般基于文档集中每两篇文章间的相似度计算,时间复杂度为 $O(n^2)$ 。对于层出不穷的网页信息,搜索引擎在进行网页收集时如果对每个新收集到的网页都在文档库中做一遍两两网页文档间的相似度比较,效率将非常低。为提高相似度处理效率,本文采用基于概念的相似度处理方法。

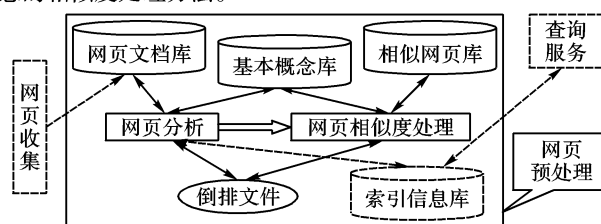


图1 系统结构

图1简要描述了搜索引擎工作的三个主要步骤:网页收集、网页预处理和查询服务。网页文档库用于存放经过网页收集获得的网页文档对象;基本概念库用于存放对网页文档对象抽象形成的概念对象;相似网页库用于存放网页相似信息的映象结构;索引信息库用于存放方便查询的网页组织结构信息。算法中网页相似度计算在预处理阶段完成。网页预处理包含了网页分析、建立索引结构等一系列内容^[9]。为清楚描述本文算法,图1仅保留与相似度计算相关的网页分析部分和网页相似度处理部分。相似度计算基于网页分析结果,依据形成的网页概念和倒排文件完成对网页的消重。

2 相似度处理模型

2.1 倒排文件结构

对于海量网页集合,建立用户检索与网页信息间的有效

联系是查询子系统工作的重点,目前最适合的数据结构是倒排文件。倒排文件是用文档包含的概念作为索引,文档作为索引目标的一种结构。通过倒排文件,在已知概念的情况下能够迅速定位目标文档达到良好的查询效果。

定义 1 一个网页文档对象 D 定义为有序对 (did, v) , 其中 did 是一篇网页文档的唯一标识, v 是一个 n 元组 $(a_1: v_1, \dots, a_n: v_n)$ 描述文档对象 D 的属性。

定义 2 一个概念对象 C 定义为三元组 (cid, c, w) , 其中 cid 是一个概念的唯一标识, c 是概念对象的基本词项, w 为 c 的量化表示形式。

创建倒排索引结构包括建立正向索引和反向索引, 图 2 建立了 4 个网页文档对象 D_1, D_2, D_3, D_4 与 7 个概念对象 $C_1, C_2, C_3, C_4, C_5, C_6, C_7$ 之间的索引关系。网页分析过程建立了网页文档对象 D 到概念对象 C 的正向索引, 如图 2(a) 所示。经过重新排序, 建立以概念对象 C 到网页文档对象 D 的反向索引, 如图 2(b) 所示。

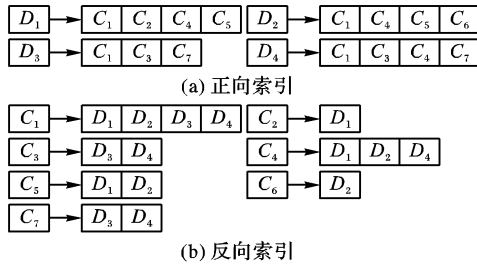


图 2 倒排文件的正向索引与反向索引

2.2 网页文档的抽象表示

在传统的文本处理领域, 一篇文档被抽象为一组对文档具有描述能力的关键词集合, 在信息检索领域有着广泛应用。其中最著名的模型是向量空间模型 (Vector Space Model, VSM)^[1,2]。对网页文档对象 D 抽象是网页分析过程的一个重要步骤, 在此过程建立从网页文档对象 D 到概念对象 C 的正向索引。

定义 3 一个网页文档对象 D 抽象表示为概念对象集合 $D = \{C_1, C_2, \dots, C_i, \dots, C_n\}$ ($1 < i < n$), 其中 C_i 为定义 2 所描述的概念对象三元组 (cid_i, c_i, w_i) , n 表示一个网页文档对象抽象形成的概念对象个数。

对网页文档对象 D 进行抽象, 形成概念对象集合 $\{C_1, C_2, \dots, C_i, \dots, C_n\}$ 需要两个步骤: 首先是网页净化, 其次是概念对象提取和量化。网页净化去掉网页中大量无用的广告、导航栏等噪音以及对概念对象形成无用的脚本标记, 以提高网页主题提炼效果和处理准确性^[3]。文献[4,5] 分别描述了一些去除噪音的方法。

在概念对象提取和量化过程中, 最重要工具是停用词。停用词是信息库的文档中出现频率过高且不具有主题区分和描述能力的功能性词语, 比如, a, an, the 等。形成概念对象 $C = (cid, c, w)$ 需要完成以下工作^[2]: 1) 识别独立单词; 2) 根据停用词表去除功能性词语; 3) 形成具有描述能力的概念对象基本词项 c ; 4) 计算基本词项 c 的量化值 w 。根据上述定义, 图 2 描述的网页文档对象被抽象为: $D_1 = \{C_1, C_2, C_4, C_5\}$, $D_2 = \{C_1, C_4, C_5, C_6\}$, $D_3 = \{C_1, C_3, C_7\}$, $D_4 = \{C_1, C_4, C_7\}$, 在此网页文档对象抽象的基础上建立正向索引。

2.3 相似度计算

相似度计算基于文档库中文档间两两距离对比完成^[7,8], 是聚类分析的一种重要方法, 在搜索引擎系统中有广泛应用。

定义 4 网页文档对象 D' 定义为需要进行相似度比较的网页文档对象 D_i 和 D_j 在概念对象基本词项上的并集。设 $D_i = \{C_{i1}, C_{i2}, \dots, C_{ik}, \dots, C_{in}\}$ ($1 < k < n$), 其中 $C_{ik} = (cid_{ik}, c_{ik}, w_{ik})$ 。 $D_j = \{C_{j1}, C_{j2}, \dots, C_{jk}, \dots, C_{jm}\}$ ($1 < k < m$), 其中 $C_{jk} = (cid_{jk}, c_{jk}, w_{jk})$ 。定义 D' 为:

$$D' = D_i \cup D_j = \{C_1', C_2', \dots, C_k', \dots, C_l'\} \quad (1 < k < l \leq n + m) \text{ 其中, } C_k' = (cid_{ik}, c_{ik}, 0), t \in \{i, j\}。$$

定义 5 网页文档对象 D_i' 定义为网页文档对象 D_i 在网页文档对象 D' 上的映射。 $D_i' = \{C_{i1}', C_{i2}', \dots, C_{ik}', \dots, C_{il}'\}$ ($1 < k < l \leq n + m$)。其中, $C_{ik}' = (cid_{ik}, c_{ik}, w_{ik}')$ 。如果 c_{ik} 是从网页文档对象 D_i 中抽取的概念对象基本词项, $w_{ik}' = w_{ik}$; 否则 $w_{ik}' = 0$ 。

定义 6 网页文档对象 D_i 和网页文档对象 D_j 间相似度 $sim(D_i, D_j)$ 为:

$$sim(D_i, D_j) = sim(D_i', D_j') = \frac{\sum_{k=1}^l w_{ik}' \cdot w_{jk}'}{\sqrt{\sum_{k=1}^l w_{ik}'^2} \cdot \sqrt{\sum_{k=1}^l w_{jk}'^2}}$$

2.4 基于概念对象的相似度处理模型

面对海量网页文档数据, 根据定义 6 描述, 如果将一个新网页文档对象 D_{new} 与网页文档库中每一个网页文档对象进行两两相似度比较, 在时间和资源上都会产生极大浪费。基于概念对象的相似度处理模型根据网页文档对象 D_{new} 的抽象表示: 概念对象集合 $D_{new} = \{C_1, C_2, \dots, C_i, \dots, C_n\}$ ($1 < i < n$, $C_i = (cid_i, c_i, w_i)$), 有选择的获取需要进行相似性验证的文档完成相似度处理。

定义 7 定义网页文档对象 D_{new} 所包含概念对象 C_i 对应的索引文档目标对象集合为 $G_i = \{D_j | C_i \rightarrow D_j, C_i \in D_{new}\}$, \rightarrow 表示在反向索引中, 存在从 C_i 到 D_j 的索引关系。

定义 8 定义对网页文档对象 D_{new} 进行相似性验证的网页文档对象集合 G 为: $G = \bigcap_{i=1}^{|D_{new}|} G_i$ 。在集合 G 的基础上进行相似度处理减少了需要计算的文档基数, 相似度计算性能得到良好改善。设网页文档对象 $D_5 = \{C_1, C_4\}$ 为待验证相似性的网页文档对象, 如图 2 所示, 根据定义 7 获得包含概念对象 C_1 对应的索引文档目标对象集合 G_1 为 $\{D_1, D_2, D_3, D_4\}$, 概念对象 C_4 对应的索引文档目标对象集合 G_4 为 $\{D_1, D_2, D_4\}$ 。根据定义 8, 与 D_5 进行相似度计算的网页文档对象集合 G 为 $\{D_1, D_2, D_4\}$ 。在 G 基础上采用定义 6 进行的相似度计算, 在效率和质量均有很好提高。

3 网页相似度判定

3.1 相似度网页库结构

相似网页库存储相似网页的聚类关系。每一个簇中的网页文档对象集合具有网页主题内容相似的特点。图 3 描述了网页文档对象 D_1, D_2, D_3, D_4 之间的相似关系。基于图 2 描述概念对象与网页文档对象间的索引关系, 定义图 3 中网页文档对象 D_1, D_2 主题相似属于簇 H_1 , 网页文档对象 D_3, D_4 主题相似属于簇 H_2 。

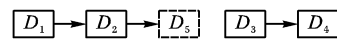


图 3 相似网页文档对象关系

根据定义 8 获取与待判断相似度网页文档对象 D_{new} 主题内容相似的网页文档对象集合 G , 对缩小待处理网页文档对象个数有很好效果。但是根据相似网页库对相似文档簇的划分, 集合 G 不会只出现在一个簇中, 因此计算相似度前应该该尽

量缩小簇的个数以更好提升计算性能。例如根据图 3 所示,与 D_5 进行相似度计算的网页文档对象集合 $G = \{D_1, D_2, D_4\}$ 中的元素分别出现在簇 $H_1 = \{D_1, D_2\}$ 和簇 $H_2 = \{D_3, D_4\}$ 中。但是 G 包含了簇 H_1 中的所有网页文档对象,而 G 只包含了簇 H_2 中的部分网页文档对象,计算相似度时只考虑 D_5 与簇 H_1 的关系。同时根据相似度计算的特点,在一个簇中无法选择具体哪一个网页文档对象对该簇更具有代表性,因此如果一个簇中满足相似度计算要求的网页文档对象个数比其他任意一个簇都大,认为网页文档对象 D_{new} 属于该簇。

定义 9 定义网页文档对象 D_{new} 相对于簇 H_j 更满足簇 H_i 的条件为: $|H_i'| > |H_j'|$, 其中 $H_i' = \{D_i | \text{sim}(D_{new}, D_i) > \beta, D_i \in H_i\}$, $H_j' = \{D_j | \text{sim}(D_{new}, D_j) > \beta, D_j \in H_j\}$ 。

3.2 相似度处理算法

在搜索引擎查询子系统中,需要对新获得的网页文档对象进行网页分析、索引建立、相似度计算等预处理,以合理组织网页文档对象和概念对象来优化查询服务。定义预处理过程如下:

```

procedure PreProcess (D Dnew)
  将网页文档对象  $D_{new}$  抽象表示为概念对象集合  $D_{new} = \{C_1, C_2, \dots, C_i, \dots, C_n\}$ 
  建立概念对象  $C_i (1 < i < n)$  到文档的反向索引
  if ( $D_{new}$  形成的概念对象大量存在于基本概念库中)
  then
    DealSimilarity ( $D_{new}$ )
  end if
end procedure

```

网页相似性验证是预处理过程的重要步骤,定义完整的网页文档对象 D_{new} 基于概念的相似度处理描述算法如下:

```

function DealSimilarity (D Dnew)
  获取网页文档对象  $D_{new}$  进行相似性验证的网页文档对象集合  $G$ 
  根据  $G$  在相似网页库中选择簇与  $D_{new}$  进行相似度计算
  将  $D_{new}$  加入合适的簇
end function

```

算法描述了进行相似度计算的完整过程,基于此算法的相似度处理能够缩小需要计算的网页文档对象范围,为能够加速整个处理过程、提高相似度计算效率起到良好的作用。

4 实验分析

该实验基于小型搜索引擎系统,实验过程采用两种相似度处理策略。方法一:对每一篇新获取的文章在数据库中进行逐条记录扫描,以进行相似度比对;方法二:采用基于概念的相似度模型进行处理。对于前者每篇文章的备选处理集是数据库中所有文档集合,后者首先形成备选处理集,在备选集的基础上进行相似度处理。实验表明,在网页文章积累初期,两种方法在耗时上并无太大区别,但是随着网页文档数目增加,后者性能明显优于前者。

表 1 不存在相似文章的记录处理

测试项目	方法一	方法二
网页产生概念个数	15	15
数据库中已存概念个数	1	1
相似度备选处理集个数	231 987	0
每 100 条比对耗时(s)	62	0
相似文章个数	0	0

部分实验数据如下:数据库现有文章记录数 231 492 条,

一次获得新网页文档记录数 830 条。进行相似度判断,对一条新记录,实验是否存在相似文章的情况,取阈值 $\text{sim}(D_i, D_j) > 0.7$, 分别比较两种方法效率。

对于新获取网页文档记录 830 条,方法二需要进行相似度处理的新文档个数为 179 条,预处理过程包含概念形成和相似度处理,耗时 2 275s。

表 2 存在相似文章的记录处理

测试项目	方法一	方法二
网页产生概念个数	15	15
数据库中已存的概念个数	12	12
相似度备选处理集个数	231 746	109
每 100 条比对耗时(s)	62	62
相似文章个数	53	51

5 结语

本文提出的基于概念对象的网页相似度处理算法,能够根据网页文档对象的分析结果迅速确定需要进行相似度处理的网页范围,为提高处理效率减少不必要的相似度计算奠定良好基础。

但是基于概念对象的相似度处理算法对概念对象精度要求比较高。随着停用词出现位置的不同,相似主题的概念也不会完全相同,处理结果会受到一定影响。因此下一步的工作重点一是放在适合相似度计算的概念对象抽取过程,形成更加精确的概念对象。二是优化计算过程涉及的概念对象选取,选择对网页文档对象更具代表性的概念对象进行计算,以提高计算结果。三是对两篇相似网页文档对象的相似度对比工作,除了依据本文提出的相似度计算公式根据概念对象进行相似度处理外,还可以使用更加精确的计算形式以提高判断计算效果。

参考文献:

- [1] SALTON G, MCGILL MJ. Introduction to Modern Information Retrieval[M]. McGraw-Hill, Inc., 1983.
- [2] SALTON G. Automatic Text Processing - the Transformation, Analysis and Retrieval of Information by Computer[M]. Addison-Wesley Publishing Co., Reading, MA, 1989.
- [3] 李晓明, 闫宏飞, 王继民. 搜索引擎——原理、技术与系统[M]. 第 1 版. 北京: 科学出版社, 2005.
- [4] SHIAN-HUA LIN, JAN-MING HO. Discovering informative content blocks from Web documents[A]. Proceedings of the SIGKDD Conference[C]. 2002. 588 - 593.
- [5] YANG YM. Noise reduction in a statistical approach to text categorization[A]. Proceedings of SIGIR295, 18th ACM International Conference on Research and Development in Information Retrieval[C]. 1995.
- [6] HAN JW, KAMBER M. Data Mining: Concepts and Techniques[M]. Morgan Kaufmann Publishers, Inc., 1998.
- [7] ETZWEILER L, MARTIN C. Binary cluster division and its application to a modified single pass clustering algorithm[R]. In Report No. ISR-21 to the National Library of Medicine, 1972.
- [8] JOON HO LEE. Combining Multiple Evidence from Different Properties of Weighting Schemes[A]. Proceeding of the 18th annual international ACM SIGIR conference on Research and development in information retrieval[C]. 1995.
- [9] BRIN S, PAGE L. The Anatomy of a Large - Scale Hypertextual Web Search Engine[A]. Proceedings of the 7th International World Wide Web Conference[C]. 1998.