面向网页分类的网页摘要方法

鲁明羽[1,2], 沈 抖[2], 郭崇慧[2,3], 陆玉昌[2]
(1. 大连海事大学计算机科学与技术学院, 辽宁大连 116026; 2 清华大学计算机科学与技术系, 北京 100084; 3 大连理工大学应用数学系, 辽宁大连 116024)

摘 要: 网页分类是网络挖掘的重要研究内容之一. 与文本分类相比, 网页分类面临的困难更多. 去除网页中的噪声信息可以提高网页分类的精度, 基于摘要的网页分类方法利用了这一思想. 本文对三种传统的网页摘要方法进行了分析和改进, 提出了 Content Body 摘要方法以及基于四种摘要方法的混合摘要方法; 在此基础上, 进行了大量基于摘要的网页分类实验. 实验结果表明, 所有的摘要方法都可以提高分类效果, 其中混和摘要方法效果最好, 可以使分类的 F1值得到 12 9%的改进.

关键词: 网页分类; 网页摘要; Content Body; 混合摘要方法
中图分类号: TP301 文献标识码: A 文章编号: 0372-2112 (2006) 08-1475-06

Web-page Summarization Methods for Web-page Classification

LU Ming-yu[1,2], SHEN Dou[2], GUO Chong-hui[2,3], LU Yu-chang[2]
(1. Institute of Computer Science and Technology, Dalian Maritime University, Dalian, Liaoning 116026, China;
2. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China;
3. Department of Applied Mathematics, Dalian University of Technology, Dalian, Liaoning 116024, China)

Abstract Web-page classification is an important research direction of web mining and much more difficult than pure-text classification. The accuracy of web-page classification can be heightened by getting rid of noisy information embedded in web pages, and the idea is utilized by our proposed summarization-based web-page classification method. In the paper, three traditional web-page summarization methods are analyzed and improved, and the Content Body summarization method and an ensemble summarization method based on four summarization methods are proposed. A large amount of experimental results of web-page classification based on summarization show that all the summarization methods can improve the performance of web-page classification algorithms and the ensemble summarization method achieves a 12 9% improvement over pure-text based methods.

Key words web-page classification; web-page summarization; content body; ensemble summarization method

1 引言

网页分类是处理海量网页信息的常用方法, 也是文本挖掘和网络挖掘的一项重要研究内容. 由于有广告栏、导航条和版权信息等噪声的存在, 网页分类比文本分类难度更大[1]. 如果直接将纯文本分类方法应用于网页分类, 则很可能受到网页噪声信息的误导和干扰, 而无法将分类的重点集中在网页主题和重要内容上, 影响分类的效果. 因此, 应设计智能的处理方法, 以便从网页中抽取主题内容.

网页摘要是指从网页中抽取关键信息, 用简洁的形式描述网页的关键 (主题)内容, 使用户不需阅读全文即可了解网页或网页集合的总体内容. 由于网页摘要的结果包含了网页中的主要结构和内容信息, 同时去除了网页中的一些噪声, 因此在网页摘要的基础上进行网页分类, 从理论上说可以提高分类精度.

我们提出的基于摘要的网页分类方法的主要思想是: 首先利用网页摘要方法滤除网页中的噪声信息, 然后在得到的网页摘要基础上进行网页分类. 为了证实网页摘要可以提高分类效果, 我们首先进行了一个理想实验, 即在人工生成的网页摘要上进行分类. 结果表明, 相对于利用网页全文进行分类, 基于摘要的分类方法可以使分类的 F1值[2,13,14]提高 14 8%. 接下来, 我们基于几种网页摘要技

收稿日期: 2005-08-16; 修回日期: 2006-01-11
基金项目: 国家自然科学基金 (No 60473115)

术进行了分类实验. 实验结果显示, 所有的摘要方法都可以不同程度地提高分类效果. 最后, 我们联合利用几个摘要方法, 提出了一个混和摘要方法, 并在分类的 F1值上取得了 12.9%的改进, 与理想实验十分接近.

2 传统摘要方法及其改进

2.1 相关工作

已经有一些研究工作尝试利用摘要来提高分类效果, 但这些工作都集中在纯文本分类领域[3~5], 简单尝试了一些基于启发式规则的摘要方法. 如文[5]把摘要当作一种特征选择的方法用来帮助分类.

与我们工作相关的另一类研究工作是分析网页结构, 过滤网页中的噪音. Yi等在文[6]中引入了一种树结构, 取名为"风格树"(Style Tree). 他们通过"风格树"来获取特定网站上的网页的表示风格并据此抽取网页的主题内容. 实验结果表明, "风格树"在分类和聚类上都可以取得明显的进步. 但是在实际应用中, 待分类的网页来自于大量不同的网站, 因此为每一个网站构造一个"风格树"是不现实的.

相对于已有工作, 我们的研究工作的意义在于: (1)将网页摘要作为一种过滤网页噪音的工具, 在网页而不是超文本上证实了摘要方法对分类的贡献; (2)在大规模数据集上证实了基于摘要的网页分类的可行性.

2.2 Luhn方法及其改进

该方法是 Luhn提出的一种基于文章表面级特征的经典摘要算法[7]. 其核心思想是为文章中的每一个句子赋予一个意义值, 那些具有最大意义值的句子将会被抽取出来作为摘要, 其中句子的意义值是通过计算句中意义词的个数得到的.

为了使 Luhn方法更好地帮助分类, 我们对其做了一点修改. 由于在分类任务中, 训练集中的网页包含类别信息, 因此"意义词"的选择可以在每一个类内进行. 首先过滤停用词, 然后选取每一个类中的高频词作为该类的"意义词集"(significant words pool), 然后再用 Luhn方法计算每一个句子的意义值.

这种改进有两点好处: (1)利用了训练集所提供的先验知识; (2)通过利用整个类内文档中词的统计信息, 可以避免仅在单个网页中频繁出现的噪音词.

此外, 如果多个用户用某一段查询文本(Query或 Query Text)来检索并访问一个网页, 可以认为这段查询文本代表了这个网页的某一方面的内容. 如果这种数据足够多, 我们有理由相信, 利用这些查询词来进行网页摘要会更加准确. 为了体现查询词在摘要中的作用, 可以利用 Web服务器的查询日志信息, 修改 Luhn算法中句子权重的计算公式, 即在统计意义词出现次数时, 提高出现在查询文本中的意义词的权重.

具体做法是: 不考虑停用词表, 仅用高频和低频两个阈值来确定"意义词". 如果查询文本中的词不在目标网页中出现, 则将其从查询文本删除. 计算剩下的查询文本中所有词在网页中出现的频度均值记为 F_{ave}, 则高频、低频阈值分别设为 αF_{ave}, 其中 α的取值可以根据实验结果进行调整, 一般可分别设为 1.2和 0.8.

2.3 LSA摘要算法及其改进

潜在语义分析(Latent Semantic Analysis, LSA)已经被成功的应用到信息检索和其他一些领域当中[8]. 它的成功源于它能够将词以及与之相关的概念表示为高维的语义空间中的点. 在文本摘要领域, Gong Yihong和 Liu Xin已经成功的将 LSA应用到文本摘要中[9].

LSA建立在奇异值分解(Singular value decomposition, SVD)之上. SVD是一个矩阵分解技术, 已经被人们大量应用到文本集上. 给定一个 m* n的矩阵, $\boldsymbol{A} = [A_1, A_2, \cdots A_n]$, 每一列 $\boldsymbol{A}_i$ 表示目标文档中一个句子的词频向量, SVD可以表示为

$$\boldsymbol{A} = \boldsymbol{U}\Sigma\boldsymbol{V}^{T} \tag{1}$$

其中 $\boldsymbol{U} = [u_{ij}]$是一个 m ×n 的列标准化正交矩阵, 它的每一列被称为左奇异向量; $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_n)$是一个 n ×n对角矩阵, 其对角线上的元素是按降序排列的非负奇异值; $\boldsymbol{V} = [v_{ij}]$是一个 n ×n 的正交矩阵, 其中的每一行被称为右奇异向量.

由于 LSA的如下两个特性, 使得它适合于摘要: 首先, LSA能够通过从语义上对词和句子进行聚类来捕捉并建模词语之间的关系; 第二, LSA能够捕获文章用来表示特定概念和主题的那些显著的、不断重现的词的组合模式. 在 LSA中, 概念是用一个奇异向量来表示的, 该向量所对应的奇异值的大小表明这个概念在文本中的重要程度, 而且任何包含这个概念的句子都可以投影到这个奇异向量上. 并且能够最好体现这个概念的句子在这个向量上的投影值最大.

在 LSA摘要算法中, 所有词都是平等的, 其权重仅由词的词频决定. 可以考虑利用某种方式(例如 TF* IDF等)对词进行加权, 从而改进 LSA方法. 同样, 我们也可以考虑利用查询日志对词进行加权, 由此可产生加权的 LSA方法 WLSA(Weighted-LSA).

所谓 WLSA是指根据查询日志中的信息, 对目标网页的"句子 – 词项"矩阵中的某些行中的元素进行加权. 这些行是指出现在查询文本中的词所在行, 加权的方式是将这些行中的元素乘以某个系数 β.

β的取值可有两种方式:

(1) β取常数, 如 2;

(2) β正比于该词在与目标网页关联的查询文本中出现的次数.

2.4 Supervised摘要方法

在实际的应用中, 如果能够获得大量的对应于网页的摘要, 那么就可以用它们作为训练集, 把文本摘要问题转化为一个分类的问题, 即把网页中的句子分为"放入摘要"

和“不放入摘要”两类, 采用分类算法来实现网页摘要.

在我们的实验中, 主要采用 8种特征进行训练, 其中有 5种是传统的文本特征, 3种是专门针对网页产生的特征. 首先定义如下一些符号表示:

PN: 段落数;

SN: 句子数;

PL: 给定段落的句子数;

Para(i): 句子 S_i 关联的段落;

TF_w: 在给定网页中词 w 出现的次数;

SF_w: 在给定网页中词 w 出现的句子数;

给定一个句子 S_i (i = 1…SN), 其 8个特征表示如下:

(1) f_{i1}: 句子在给定段落中的位置信息. 它的计算公式是: Max(1/i, 1/(PL-i));

(2) f_{i2}: S_i 的长度, 例如 S_i 中所包含的词个数. 根据观察, 太长或太短的句子都不适合放入摘要. 因此设置句子最短不少于 5个单词, 最长不超过 35个单词. 在这中间的设 f_{i2} = 1, 否则 f_{i2} = 0;

(3) f_{i3}: $\Sigma TF_w * SF_w$. 这个特征不仅考虑到词的出现频率, 还考虑到词的分布特征.

(4) f_{i4}: S_i 与标题间的相似度. 相似度是通过计算它们之间的点积得到的.

(5) f_{i5}: S_i 与目标网页中所有的文本间的相似度;

(6) f_{i6}: S_i 与 meta数据间的相似度;

(7) f_{i7}: S_i 中特定单词的出现频率. 这些特定单词是指网页中具有斜体、加粗、下划线等表示强调的单词集中的词.

(8) f_{i8}: 表示 S_i 中平均字体大小. 在网页中, 通常越大的字体表示的内容越重要. 从网页中抽取出上述 8个特征后, 就可以利用朴素贝叶斯分类器实现摘要.

对每一个句子 s, 根据贝叶斯先验概率, 判断它是否包含在摘要 S 中. 根据给定的 k 个特征: F_j, j= 1…k, 计算它的概率, 用贝叶斯规则表示为:

$$P(s\in S \mid F_1, F_2, \ldots F_k) = \frac{P(F_1, F_2, \ldots, F_k \mid s\in S)P(s\in S)}{P(F_1, F_2, \ldots F_k)}.$$

假定统计特征相互独立:

$$P(s\in S \mid F_1, F_2, \ldots F_k) = \frac{\prod_{j=1}^{k} P(F_j \mid s\in S)P(s\in S))}{\prod_{j=1}^{k} P(F_j)} \quad (2)$$

$P(s\in S)$是一个常量, 即压缩率, $P(F_j \mid s\in S)$和$P(F_j)$可以直接根据训练文档集中的出现词数进行估计.

计算出每个句子属于摘要的概率后, 可以按照概率的大小对句子进行排序, 然后从高到低选出一定数目 (即满足压缩率要求)的句子作为摘要.

3　Content Body摘要算法和混合摘要方法

3.1　Content Body摘要算法

我们从网页的结构特征出发, 通过分析网页的结构和层次特征, 提出了一种基于网页分割技术的 Content Body 摘要算法(简称为 CB). 该方法将网页分割成不同的对象, 然后分析各个对象之间的相似关系, 进而提取出网页的主体内容.

对于网页摘要来说, 所应提取的是网页主题内容中最重要的信息, 而网页除了主题内容之外, 还含有大量的无关信息, 包括多媒体信息、导航信息和广告、版权和地址信息等其他信息. 我们把网页设计者为了辅助网站组织而增加的信息定义为“噪声”, 把原本要表达的核心文字素材称为“主题内容”. 当噪声和主题内容已经混合在一个半结构化的 HTML 文件中时, 想依据设计者的意图把两者分开并不是一件容易的事情. 我们采用功能对象模型 (Function-Based Object Model, FOM)[10]的一个简化版本来解决这个问题.

FOM 模型通过分析网页中每一个对象的功能和类别来理解网页设计者的意图. 在 FOM 模型中, 有两种粒度的对象, 一种是基本对象 (Basic Object, BO), 一种是复合对象 (Composite Object, CO). 基本对象是指网页中不能被分割的最小的信息单位; 复合对象是指用来表示特定功能的一组基本对象或复合对象的集合. 在 HTML 表示的网页中, 一个基本对象是指包含在两个标签 (tag)之间的不可分割的元素, 也就是说, 这个元素中不再包含其他的标签. 根据这个定义, 很容易找到网页中所有的基本对象. 同样, 可以通过分析网页的布局找出复合对象. 这里用的一个假设是指, 属于同一类的对象通常具有一致的视觉风格, 因此不同类别的对象之间会有明显的视觉边界.

找出网页中所有的基本对象和复合对象后, 可以根据一些启发式规则来识别这些对象的类别. 这里用到的一些对象类型包括:

(1)信息对象 (Information Object), 用来表示内容信息;

(2)导航对象 (Navigation Object), 用来提供导航信息;

(3)交互对象 (Interaction Object), 用来提供用户和网站的交互;

(4)修饰对象 (Decoration Object), 用来修饰网页;

(5)其他特定功能对象 (Special Function Object), 用来实现某种特定的功能, 如广告、版权、Logo等.

通过分析 FOM 模型中各个对象之间的关系, 可以用来帮助生成网页摘要

CB摘要方法的思路是依据一定的相似度量方法, 在功能对象模型下从一个网页中所有的对象中选出一个核心对象 (Core Object), 然后把与核心对象有密切关系的对象选出, 共同组成该网页的摘要.

具体方法如下:

(1)根据 FOM 模型找出一个网页中所包含的对象, 将明显不是网页主要内容的对象滤除, 如修饰对象、广告对象等;

(2)将过滤后的每一个对象当作一篇文档,构建它们的 TF* IDF索引;

(3)计算其中任意两个对象之间的相似度,相似度计算方法为余弦相似度.如果两者之间的相似度超过一个阈值,则在两者之间添加一条边.阈值是根据经验来选择.通过该步计算后可以在网页的对象上构建一个图;

(4)定义"核心对象"(Core Object)为图中相似度最大的对象;

(5)定义"主题对象"为核心对象以及与核心对象之间有边连接的对象.主题对象的内容——"主题内容"即是网页的摘要结果.

虽然上述 Content Body摘要算法可以有效地滤除网页中的噪音,但是存在几个不足之处:

(1)生成摘要没有考虑冗余;

(2)在多主题的网页中,难以覆盖所有的主题;

(3)含有复杂对象的单主题网页的摘要效果不理想.

基于查询日志,我们可以对上述方法进行改进,产生新的选择核心对象和"主题内容"的方法.

假设网页 P 中共有 m 个对象,对象 O_i ($i=1, 2, \cdots, m$)的"度"为 D_i.另设查询文本中所有的词在 P 中出现的次数为 n,在 O_i 中出现的次数为 F_i,则可以按下式赋予对象 O_i 一个权重 W_i:

$$W_i = D_i/m + F_i/n \tag{3}$$

计算出每个对象的权重之后可以按下式选取核心对象

$$\text{Core Object} = \arg\max_{O_i}(W_i) \tag{4}$$

选出核心对象以后,可以如前所述选出"主题内容",完成网页摘要.当然也可以直接选择 W_i 值最大的 k 个对象,由它们组成网页的摘要,k 由摘要压缩率决定.

3.2 混和摘要方法

该方法是在 Luhn 方法、LSA、Content Body 以及 Supervised等四种网页摘要方法基础上,对这四种摘要方法进行组合,产生的一个新的混合摘要方法.

为了解释混和摘要方法,首先定义几个表示符: S_{luhn}, S_{lsa}, S_{cb}, S_{sup}.它们分别表示在网页中的某个句子在 Luhn、LSA、Content Body、Supervised摘要方法中获得的权重.其中 S_{luhn}即为该句子的意义值; S_{lsa}指该句子在其所对应的奇异向量上的索引值; S_{cb}取值为 0和 1,如果该句子属于"主题内容"则为 1,反之为 0, S_{sup}是指该句子属于摘要的概率,由摘要器即分类器给出.

给定一个网页,分别利用 Luhn、LSA、Content Body、Supervised四种方法计算出各个句子的权重,然后把四个权重按一定比例相加,即可得到该句子的一个混和权重,然后选择混和权重最高的句子作为该网页的摘要,这就是混和摘要方法.混和摘要方法中句子权重的计算公式如下:

$$S = w_1 S_{luhn} + w_2 S_{lsa} + w_3 S_{cb} + w_4 S_{sup} \tag{5}$$

4 网页分类实验与结果分析

为了检验所提出的基于摘要的网页分类方法的有效性,我们作了一系列对比实验.首先,在人工产生的网页摘要基础上进行分类,以验证网页摘要确实有助于网页分类;其次,重点比较所提出的 Content Body摘要方法与改进的 Luhn方法和 LSA 方法以及 Supervised摘要方法的分类效果;最后,评估混合摘要方法的分类效果.

实验中我们还研究了不同的参数设置对于摘要效果的影响,限于篇幅,未及详述.

4.1 网页分类器与实验数据

由于我们的目标是重点检验网页摘要对于网页分类的有效性,因此,我们没有选择过于复杂的分类器,而是直接采用我们研制的 SCETCS系统中的朴素贝叶斯分类器和支持向量机分类器[11].评价方法以基于网页中的纯文本内容的分类作为底线(Baseline).其中,采用的 SVM 是 J Platt给出的序列最小优化算法(Sequential Minimal Optimization, SMO)的一个简单快速版本 SMOX[12].

实验用的网页数据采用从 LookSmart站点(http://search.looksmart.com)下载的 2百万个网页,其中 50万个网页包含了人工摘要.这里的人工摘要是指网页的作者给出的关于网页主题内容的一个描述,这些描述不一定是完全对应于网页中的句子.由于在 50万个网页的大规模数据上进行实验比较耗时,我们随机选择了其中 30% 的网页用于实验.随机选出的数据集包括 153,019个网页,这些网页分布在 64个类中(这里我们只考虑 LookSmart网站上最上两层的类别).其中最大的类"Library\Society"包括 17,473个网页,最小的类"People & Chat\Find People"包括 52个网页.

表 1 三个最大的类

Category Name	Total	Train	Test
Library\Society	17473	15726	1747
Travel\Destinations	13324	11992	1332
Entertainment\Celebrities	10112	9101	1011

表 2 三个最小的类

Category Name	Total	Train	Test
Sports\News & Scores	106	96	10
People & Chat\Personals	74	67	7
People & Chat\Find People	52	47	5

表 1和表 2分别给出了最大和最小的三个类的情况.为了降低随机性,我们采用 10-fold交叉验证方法[13].

4.2 实验结果

4.2.1 底线实验

最简单的网页分类是把网页看作纯文本.这种方法首先滤掉网页中的 HTML标签,将网页转换为纯文本;其次通过滤除停用词和求词根(stemming)将网页表示成一个词袋(Bag-of-words),每一个单词的权重用它们在网页出

现的次数(即 TF)表示. 在很多实验中, 研究人员把这种方法作为底线. 在我们的实验中, 同样采用这种方法作为底线, 实验结果如表 3和表 4 Full-text行所示.

从表中可以发现, SMOX 的分类结果为 0. 651(micro-F1), 相对于 NB算法高出 2. 4%. 同时还可以发现, 在 10-fold上的方差仅为 0. 3%, 说明在这个数据集上的分类效果是稳定的.

表 3　NB算法上的实验结果

method	microP	microR	micro-F1
Full-text	70. 7±0. 3	57. 7±0. 3	63. 6±0. 3
Title	68. 3±0. 4	55. 4±0. 4	61. 2±0. 4
Meta-data	47. 7±0. 4	38. 7±0. 4	42. 7±0. 4
Description	81. 5±0. 4	66. 2±0. 4	73. 0±0. 4
Content Body	77. 2±0. 4	62. 7±0. 4	69. 2±0. 4
Luhn	77. 9±0. 4	63. 3±0. 4	69. 8±0. 5
LSA	75. 9±0. 4	61. 7±0. 4	68. 1±0. 5
Syoervused	75. 2±0. 4	60. 9±0. 4	67. 3±0. 4
Hybrid	80. 2±0. 3	65. 0±0. 3	71. 8±0. 3

表 4　SMOX 算法上的实验结果

method	microP	microR	micro-F1
Full-text	72. 4±0. 3	59. 3±0. 3	65. 1±0. 3
Title	68. 8±0. 3	55. 9±0. 3	61. 7±0. 3
Meta-data	47. 8±0. 4	38. 8±0. 4	42. 8±0. 4
Description	82. 1±0. 4	66. 9±0. 4	73. 7±0. 4
Content Body	78. 6±0. 3	63. 7±0. 3	70. 3±0. 3
Luhn	77. 3±0. 3	62. 8±0. 3	69. 3±0. 3
LSA	79. 2±0. 3	64. 3±0. 3	71. 0±0. 3
Supervised	76. 3±0. 4	61. 8±0. 4	68. 3±0. 4
Hybrid	81. 1±0. 3	65. 7±0. 3	72. 6±0. 3

4. 2. 2　人工摘要的实验结果

为了测试摘要技术对网页分类的贡献, 我们做了一个可行性研究实验, 把网页作者关于网页主题内容的描述作为网页的“理想”摘要, 并用它们代替网页的全文进行分类. 这个实验可以帮助我们发现, 在最理想的情况下, 摘要是否能提高分类的性能. 此外, Title和 Meta数据在某种程度上也可以看作是人工摘要, 因此我们也对它们的性能进行了测试.

表 3, 表 4中 Description, Title和 Meta-Data各行分别表示在人工摘要、标题和元数据上的分类结果. 相对基于网页全文的分类效果相比, 基于人工摘要的分类算法在两个分类器上都取得了 13. 2%以上的提高, 而基于标题和元数据的分类效果反而降低了.

通过分析数据可以发现, 人工摘要确实可以帮助用户很容易的了解网页的内容, 或者说, 人工摘要较准确地概括了网页的主题. 尽管标题和元数据在某种程度上也能达到这种效果, 但经分析可知, 它们很难完整代替网页本身.

通过上述实验可以清楚地看出, 理想的网页摘要确实可以提高分类的效果, 但差的摘要反而会降低分类效果, 因此合适的摘要算法成为改进分类效果的关键.

4. 2. 3　非监督(unsupervised)摘要方法的实验结果

在非监督摘要方法中, 考查了基于网页的结构信息抽取网页主题内容的 Content Body算法以及改进的 Luhn和 LSA 算法. 我们需要确定一个阈值来判断网页中的两个对象之间是否需要添加一条边. 实验中该阈值设为 0. 1, 此外, Luhn和 LSA 两种算法的压缩率分别设为 20%和 30%.

表 3, 表 4给出了这三种摘要方法上的实验结果. 结果显示, 这三种摘要方法对分类的贡献相当, 它们相对底线都取得了 7%以上的提高.

4. 2. 4　supervised摘要方法的实验结果

由于本实验中的人工摘要是网页的作者给出的关于网页主题的描述, 而不是直接从网页中抽取得到的, 它们不能直接用作 supervised摘要方法所需要的训练数据. 因此, 需要根据一定的原则, 从中生成训练数据. 在本实验中, 如果一个句子跟人工摘要的相似度超过一个阈值(0. 3), 这个句子被标为正例(即应该放入摘要), 反之标为负例. 基于 supervised摘要方法的实验结果如表 3, 表 4所示, 表中数据是压缩率在 20%时的结果.

基于有指导摘要方法的分类结果相对于底线提高了 6%, 低于基于非监督摘要方法. 其原因可能是由于训练数据不够准确, 因为是根据它们与人工摘要的相似性来确定的.

4. 2. 5　混和摘要方法的实验结果

通过上述实验可以发现, 无论是 supervised摘要方法还是非监督摘要方法都可以在一定程度上提高分类效果. 但它们都与基于人工摘要的分类效果有较大差距. 混和摘要方法实验通过组合这些摘要方法, 试图做到取长补短, 进一步提高分类的效果.

从表 3, 表 4可以发现, 混和摘要方法相对于底线取得了 12. 9%的进步, 逼近了基于人工摘要的进步. 这里的混和摘要算法中各个摘要方法的权重相同.

5　结语

本文的摘要算法把每一个网页作为独立的文档来处理, 然而万维网上的网页之间存在着丰富的关联关系, 如果能够充分利用这些关系, 应该能够进一步提高摘要效果, 进而提高分类的效果. 此外, 在今后的工作中, 希望用统一的框架来描述本文的思想, 例如拟运用 Boosting机制, 对 SVM 分类器进行训练, 提高基于摘要的网页分类方法的性能.

参考文献:

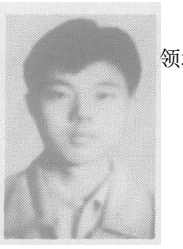
[1] Z Chen, S P Liu, W Y Liu, G G Pu, W Y Ma. Building a web thesaurus from web link structure[A]. Proc of the 26th annual international ACM Sigir[C]. Toronto, Canada: ACM Press, 2003. 48-55.

[2] D Shen, Z Chen, Q Yang, H J Zeng, B Y Zhang, et al. Web-

page classification through summarization[A]. Proc of the 27th Annual International ACM SIGIR[C]. Sheffield, UK: ACM Press, 2004. 242-249.
[3] S J Ker, J N Chen. A text categorization based on summarization technique[A]. Proc of the 38th Annual Meeting of the Association for Computational Linguistics IR&NLP workshop[C]. Hong Kong, China: Association for Computational Linguistics, 2000. 79-83.
[4] Ko, J W Park, J Y Seo. Automatic text categorization using the importance of sentences[A]. Proc of the 19th International Conference on Computational Linguistics[C]. Taipei: Morgan Kauffmann Press, 2002. 1-7.
[5] A Kolcz, V Prabakarmurthi, J K Kalita. Summarization as feature selection for text categorization[A]. Proc of the 11th International Conference on Information and Knowledge Management[C]. Atlanta, USA: ACM Press, 2001. 365-370.
[6] L Yi, B Liu, X Li. Eliminating noisy information in web pages for data mining[A]. Proc of KDD2003[C]. Washington, USA: ACM Press, 2003. 296-305.
[7] H P Luhn. The automatic creation of literature abstracts [J]. IBM Journal of Research and Development, 1958, 2(2): 159-165.
[8] J T Sun, Z Chen, H J Zeng, Y C Lu, C Y Shi, W Y Ma. Supervised latent semantic indexing for document categorization[A]. Proc of the Fourth IEEE International Conference on Data Mining[C]. Brighton, UK: IEEE Computer Society, 2004. 535-538.
[9] Y H Gong, X Liu. Generic text summarization using relevance measure and latent semantic analysis[A]. Proc of the 24th annual international ACM SIGIR[C]. New Orleans, USA: ACM Press, 2001. 19-25.
[10] J L Chen, B Y Zhou, J Shi, H J Zhang, Q F Wu. Function-based object model towards website adaptation[A]. Proc of WWW 10[C]. Hong Kong, China: ACM Press, 2001. 587-596.
[11] M Y Lu, K Y Hu, Y W u, Y C Lu, L Z Zhou. SCETCS: towards improving VSM and naive bayes classifier[A]. Proc of 2th IEEE International Conference on Systems, Man and Cybernetics[C]. Hammamet, Tunisia: IEEE Computer Society, 2002. 465-469.
[12] Platt. Sequential Minimal Optimization[DB/OL]. http://research.microsoft.com/~jplatt/smo.html.
[13] 沈抖. 万维网上数据处理方法的研究[D]. 北京: 清华大学计算机系, 硕士论文, 2004.
[14] 鲁明羽. 数据挖掘和网络挖掘若干方法及应用研究[R]. 北京: 清华大学计算机系, 博士后出站研究报告, 2005.

作者简介:

鲁明羽 男, 1963年出生, 教授, 博士生导师, 主要研究领域为数据挖掘、文本挖掘、网络挖掘、机器学习.
E-mail: lumingyu@tsinghua.org.cn



沈 抖 男, 1979年出生, 博士生, 主要研究领域为文本挖掘、网络挖掘.

郭崇慧 男, 1973年出生, 副教授, 主要研究领域为数据挖掘、机器学习、优化理论与方法、经济系统分析.

陆玉昌 男, 1937年出生, 教授, 主要研究领域为人工智能、数据挖掘、机器学习.