文章编号:1001 - 9081(2006)03 - 0638 - 03

基于 Corpus 库的词语相似度计算方法

章志凌,虞立群,陈奕秋,罗海飞,邵晓敏
(上海交通大学 软件学院,上海 200030)
(steven_zzl@263.net)

摘 要:构建了一个语义关联库,称为 Corpus 库,该库使用词语空间和关系空间结构化地存储了词语和其上下文之间的统计信息,并通过阅读大量的预料数据来训练其相关数据。详细介绍了 Corpus 库的训练方法,并对训练过程中出现的大量关系提出了裁剪方案。在此基础上,通过构建词语的上下文关系向量提出了一种词语相似度算法。实验证明这是一种有效的对词语相似度进行计算的方法。

关键词:Corpus;词语相似度;信息检索
中图分类号: TP391.1; TP18 **文献标识码**:A

Measurement of word similarity based on Corpus

ZHANG Zhi-ling, YU Li-qun, CHEN Yi-qiu, LUO Hai-fei, SHAO Xiao-min
(*Software College, Shanghai Jiao Tong University, Shanghai* 200030, *China*)

Abstract: A semantic relevant database named Corpus was built to store the required information in word similarity measurement. Corpus got the information from large scale text training and store the information in word space and relation space after analysis and tailoring. The word similarity measurement algorithm by constructing the context relation vectors based on Corpus was given, which proved to be a feasible method by experiments.

Key words: Corpus; word similarity; information retrieval

0 引言

自然语言的词语之间有着非常复杂的关系,在实际的应用中,有时需要把这种复杂的关系用一种简单的数量来度量,而词语相似度的量化就是其中的一种。词义相似度计算在信息检索、信息抽取、文本分类、词义排歧、基于实例的机器翻译等方面有广泛的应用。

本文的研究背景是基于用户爱好的智能电视节目选择。在智能电视节目选择中,当节目推荐引擎得到用户的检索词或已有用户爱好的关键词时,需要从大量的节目文本摘要信息中进行检索,并选出最符合用户口味的节目内容,然后推荐给用户。在这个检索过程中如果只使用关键词匹配技术往往会遇到词汇不匹配的问题。为了更好地基于用户的爱好信息进行节目的智能推荐,必须量化词和词之间的相似度,从而对相似的词语一并进行检索;同时要求这种相似度的度量又是可以通过学习现实世界的文本而不断动态调整的。

1 Corpus 库的结构

在计算词和词的相似度时,需要得到一些相关的信息。具体需要哪些信息取决于词和词之间的相似度量化算法。词语相似度的计算方法一般分为两种,一种是根据某种世界知识或分类体系来计算,另一种利用大规模的语料库进行统计。

在前一种计算方法中,量化所需的相关信息一般来自于某个已有的语义词典库本身。基于《知网》以及基于 WordNet 进行词语相似度量化的计算都是直接从语义词典获取信息,由于这种语义词典是经过专家精心设计的,所以能够使用较小的空间来反映词和词之间错综复杂的关系。但缺陷是词和词之间的关系是固定不变的,因此既无法根据现实世界的变化调整词和词之间的关系,也很难引入新的名词。

在后一种计算方法中,词和词的关系是通过从大规模语料的学习中得来的词和词在上下文中的共现信息,文献[1]用 Hopfeild 神经网络进行词和词的联想,并用一个反映关键词之间的关联度的模糊自反矩阵来存储词和词之间的相似度量值。对于一个有 N 个词汇的词空间 Ω,其词语关系的存储空间复杂度将达到 $O(N^2)$。同时词对相似度的计算所需信息是直接来源于语料训练库的,自反矩阵中存储的是计算后词对相似值,实际上丢失了相关的词频、词距等历史信息,很难进一步扩展和学习。

本文将介绍一种优化了的 Corpus 库,其目的是把在大规模语料库中统计得来的丰富信息进行筛选并存储,作为以后词和词之间相似度量化的信息基础。Corpus 库用于把浩瀚的语料库中所蕴含的词和词之间的关系通过统计的方法提取出来并进行存储,然后为上层的词语关系量化计算提供支持。

我们把 Corpus 库 C 定义为 $C = (W,R)$。W 表示词汇空

收稿日期:2005 - 09 - 30 **修订日期**:2005 - 12 - 21
基金项目:交大数字家电实验室"Advanced information retrieval technology using the knowledge base"项目
作者简介:章志凌(1981 -),男,上海人,硕士研究生,主要研究方向:智能信息推荐; 虞立群(1964 -),男,浙江慈溪人,教授,双硕士,主要研究方向:家庭媒体网络; 陈奕秋(1964 -),男,浙江缙云人,副教授,双硕士,主要研究方向:数字家电; 罗海飞(1964 -),男,湖北武汉人,硕士研究生,主要研究方向:文本分类; 邵晓敏(1964 -),男,江苏南京人,硕士研究生,主要研究方向:信息检索.

间,$W = \{w_1, w_2, \cdots, w_n\}$,向量 $w_i = (wname, wfreq)$; R 表示关系空间,$R = \{r_1, r_2, \cdots, r_n\}$,向量 $r_k = (wscr, wdes, cofreq, counfreq, \overline{codist}, \overline{cointerval_m})$,各分量的含义如表 1。

表 1　Corpus 库结构

空间	分量	含义
词汇空间	*wname*	词汇
	wfreq	词汇出现次数
关系空间	*wscr*	关系源,对应 W 中的一个词汇
	wdes	关系目的,对应 W 中的一个词汇
	cofreq	词对共现次数
	counfreq	词对最近一次非共现次数
	$\overline{codist}$	词对共现平均距离
	$\overline{cointerval_m}$	词对最后 m 次共现平均间隔

对于分量的选择我们遵循两个原则:1)能够以最小代价提供词语相似度计算所需的所有信息; 2)能够灵活进行信息的动态调整和关系裁剪。

这里,词汇空间 W 用于存储孤立词汇的信息,而关系空间则用于存储词汇之间的丰富的共现信息。显然,关系空间把词汇空间的向量连成了一个有向的网络(如图 2),这和现实世界中词和词之间错综复杂的关系是相对应的。

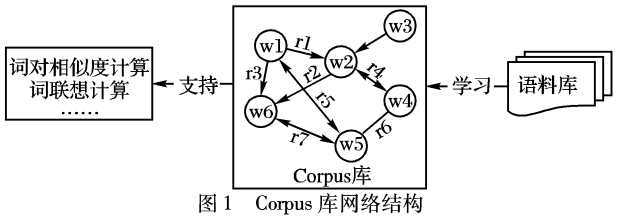

图 1　Corpus 库网络结构

同时,为了约束关系空间 R 的增长速度,我们引入容积常数 k,使得 k 满足:

$$k \approx \sum_{w_i \in W} k_i / n \tag{1}$$

这里 k_i 指以节点 w_i 为 *scr* 的关系 r 的数目,n 指词汇空间 W 所有的节点总数。当 k 较小时,Corpus 库的空间复杂度为 $O(n)$。

2　Corpus 库的构建方法

基于以上结构,我们将通过学习大规模语料来构建 Corpus 库的内容。同时由于对于给定的容积常数 k,必须使(1)式成立,因此必须动态选取最有价值的关系,并删除多余的关系。

2.1　词汇空间的构建

为了缩小 Corpus 库的容量,词汇空间 W 中的向量将只涉及二字及二字以上的名词,考虑到在电视节目的检索中,不管是新闻、体育还是文艺,地名和人名都占有相当的份额,且对检索结果有一定影响,比如“篮球”和“姚明”,“海啸”和“印尼”等,因此词汇向量除了普通名词外,还可以是地名和人名。

词汇空间 W 的构建算法可以简要描述如下:

算法 1(初始:$W = \varnothing$,任一时刻,对于输入的语料中的一个名词,若该词已在 W 中,则 *wfreq* 加 1;否则,把该词添加进 W 中,*wfreq* 设为 1)。

2.2　关系空间的构建

关系空间 R 用于记录词和词之间的共现关系。

首先对共现进行定义,我们首先引入窗口常数 $\overline{\omega}$,对语料 L 中词 w_i 的某次出现,对其前后 $\overline{\omega}$ 个词进行观察,并得到词集 $W_{\omega}^{-} = \{w_1, w_2, \cdots, w_{\omega}^{-}\}$,若发现词 $w_j \in W_{\omega}^{-}$,则说词 w_j 对词 w_i 在窗口 $\overline{\omega}$ 中共现。

$\overline{codist}$ 记录词对共现平均距离:

$$\overline{codist} = \sum_{k=0}^{cofreq} codist_k / cofreq \tag{2}$$

其中 $codist_k$ 表示词 *wscr* 和词 *wdest* 某次共现时两词中所隔的名词数量。

$\overline{cointerval_m}$ 记录词对最后 m 次共现平均间隔:

$$\overline{cointerval_m} = \frac{wfreq_m}{m} \tag{3}$$

其中 $wfreq_m$ 表示最近 m 次中词 *wscr* 新增的出现的次数。这里 $\overline{cointerval_m}$ 实际上记录了词对共现的动态数据,且 m 的大小将决定这种动态信息的更新度。

关系空间 R 的构建算法可简单描述如下:

算法 2(初始:$R = \varnothing$,任意时刻,观察到词 w_i 和 w_j 共现,若在 R 中,描述 $w_i \rightarrow w_j$ 的关系向量 r_{ij} 已经存在,则更新向量 r_{ij} 的其他分量;否则,加入新关系 r_{ij} 到 R 中,并为其他分量设定初值。)

2.3　关系空间的裁剪

事实上,算法 2 只考虑了关系空间 R 构建方法的一部分,即如何不断学习新的关系。为了满足式(1)提出的容积常数 k,需要不断删减已有的关系来限制关系空间的增长,这里涉及两个问题,一是对于以某个特定的词 w_i 为 *wscr* 的关系向量 r 的数量上限 k_i 如何确定,二是对于以某个特定的词 w_i 为 *scr* 的关系向量 r 的数量超过 k_i 时,如何进行筛选。

2.3.1　基本裁剪算法

先考虑第一个问题,我们使用如下公式确定 k_i:

$$k_i = \frac{\lg wfreq_i}{\lg \overline{wfreq}} \times k \tag{4}$$

其中 $\overline{wfreq}$ 表示词汇空间 W 中所有词语的平均词频,$wfreq_i$ 表示词 w_i 的词频。这里,词频较高的词语将享有较多的直接关联节点。同时,可以证明,k_i 的算术平均满足公式(1)。

确定了 k_i 后,现在考虑第二个问题,即如何从由 w_i 引出的所有 n_{w_i} 个关系中筛选出 k_i 个关系。这里使用共现关系权重 $weight(r)$ 作为关系重要程度的度量。对从 w_i 到 w_j 的关系 r,有:

$$weight(r) = freq_{w_iw_j} \times \overline{rel_{w_iw_j}} \tag{5}$$

其中 $freq_{w_iw_j}$ 表示 w_i 到 w_j 的共现次数,$\overline{rel_{w_iw_j}}$ 表示 w_iw_j 共现时的平均相关度:

$$\overline{rel_{w_iw_j}} = \frac{\alpha}{\overline{dist_{w_iw_j}} + \alpha} \tag{6}$$

$\overline{dist_{w_iw_j}}$ 表示 w_iw_i 共现时的平均距离,α 为可调距离常数,表示 $\overline{rel_{w_iw_j}} = 0.5$ 时,$\overline{dist_{w_iw_j}}$ 的应取值。易证,$0 \leqslant \overline{rel_{w_iw_j}} \leqslant 1$。

关系空间 R 的裁剪算法可简单描述如下:

算法3(对于 W 中每一个词汇 w_i,计算 w_i 所对应的关系上限 k_i,对以 w_i 为 *wscr* 的关系集 R_i,计算每个 $r \in R'$ 的共现关系权重 $weight(r)$,保留 R_i 中 $weight(r)$ 最大的 k_i 个关系,删除其余关系。)

2.3.2 改进的裁剪策略

在实际学习过程中,新关系往往没有足够的时间增大其共现关系权重 $weight(r)$,从而无法准确反映现实社会词语的动态变化。这里讨论对裁剪策略进行改良。

首先引入遗忘因子 $T_{mw_iw_j}$,记录 w_i 和 w_j 近期 m 次的共现变化率:

$$T_{mw_iw_j} = \lg \frac{freq_{w_i}}{freq_{w_iw_j}} / \lg interval_{mw_iw_j} \quad (7)$$

即历史共现间隔和近 m 次的共现间隔的对数比。其中 $interval_{mw_iw_j}$ 表示近 m 次共现的间隔,该值用于动态的反映 w_i 和 w_j 近期的共现频率变化情况。改良后的 $weight(r)' = weight(r) \times T_{mw_iw_j}$,可见近期拥有较高共现率词对,将得到更大的关系权重加权,并更有可能得到保留。

同时,考虑关系空间裁剪的时机,引入缓冲深度参数 $\delta(\delta > 1)$,令 Corpus 库只在子关系空间 R_i 的元素数量超过 $\delta \times k_i$ 时才进行裁剪动作。δ 使得关系子空间的两次裁剪之间有充分的缓冲区能够让近期共现关系权重较大的关系保留下来,同时 δ 也增加了空间开销。

3 词语相似度量化算法

Corpus 库可看作以词汇为节点的有向关系网络,通过对大规模语料的学习,对任一个节点词汇 w_i,该网络既存储了 w_i 自身的信息,如词频等,也有选择地存储了和 w_i 共现权重最高的 n 个词的共现信息,包括共现次数、平均共现距离等,其中 n 的选择和该词的词频成正比。通过 Corpus 库所存储的共现关系网络,每个词可以马上关联出 n 个直接共现词,再通过这 n 个词关联出大约 n^2 个二级共现词,并以此类推,形成一棵以该词为根的生成树。对于词汇 w_i 的直接共现的词 w_j,可以计算其共现权重 $weight_{w_iw_j}$:

$$weight_{w_iw_j} = \frac{freq_{w_iw_j}}{freq_{w_i}} \times \frac{\alpha}{\overline{dist_{w_iw_j}} + \alpha} \quad (8)$$

其中 $\frac{freq_{w_iw_j}}{freq_{w_i}}$ 表示 w_i 对于 w_j 的共现频率,$\overline{dist_{w_iw_j}}$ 表示 w_i w_i 共现时的平均距离,α 为一可调常数,这里 $0 \leqslant weight_{w_iw_i} \leqslant 1$。以此推论,对于词汇 w_i 在其生成树的 γ 层上共现的词 w_k,用 w_i 到 w_k 的通路中各级直接共现词权重的乘积来计算 $weight_{w_iw_k}$:

$$weight_{w_iw_k} = \prod_{w_iw_m \in V_{ik}} weight_{w_iw_j} \quad (9)$$

对任意词 w_i,可以抽取其生成树前 κ 层的 N 个共现词作为 w_i 的 N 个特征项,从而可以得到 w_i 的 N 维上下文特征向量 $d_i = (we_{i1}, we_{i2}, \cdots, we_{in})$。其中 we_{ij} 表示词 w_i 在 j 维的权重,可直接由(8)式计算。然后使用余弦系数计算词对 w_i 和 w_j 的相似度 $sim(w_i, w_j)$:

$$sim(w_i, w_j) = sim(d_i, d_j) = \frac{\sum_{k=1}^{M} we_{ik} \times we_{jk}}{\sqrt{\left(\sum_{k=1}^{m} we_{ik}^2\right)\left(\sum_{k=1}^{m} we_{jk}^2\right)}} \quad (10)$$

其中,d_i,d_j 分别为 w_i, w_j 的特征向量,M 为对齐后的特征向量的维数,we_k 为向量的第 k 维。

4 实验结果

4.1 Corpus 关系空间实验

表 2 Corpus 库训练参数

常数	含义	取值
$\overline{\omega}$	窗口常数	30
α	距离常数	10
k	容积常数	15
δ	缓冲深度常数	3

本实验考察通过 Corpus 库中词和与该词相关的关系空间中的词语集的权重,这是使用 Corpus 进行词对相似度比较的基础,其计算公式见(5)。共选择 6048 篇文章对 Corpus 库进行训练,训练过程中,使用各常数如表 2。

训练后选取词“篮球”、“运动”、“宗教”和“电脑”,其关系空间中词语的相关关系空间如表 3。

表 3 Corpus 库词语权重观察

篮球		运动		宗教		电脑	
词语	权重	词语	权重	词语	权重	词语	权重
运动	0.21	运动员	0.43	文化	0.10	网络	0.10
教学	0.13	篮球	0.23	艺术	0.09	技术	0.08
运动员	0.12	受伤	0.21	民族	0.09	信息	0.06
足球	0.10	体力	0.18	信仰	0.09	家庭	0.05
水平	0.09	健美	0.13	政治	0.08	家用	0.05
技术	0.08	技术	0.12	社会	0.06	个人	0.04
中国	0.07	足球	0.12	改革	0.03	软件	0.04
我国	0.06	全民	0.06	方式	0.03	系统	0.04
考试	0.06	体育	0.06	世界	0.03	专业	0.03
模式	0.06	水平	0.04	城市	0.03	设备	0.03
排球	0.01	耐力	0.03				

可以看出,每个词语的相关关系空间基本表征出了该词的上下文环境,事实上该词的上下文空间可以通过该词的生成树和式(9)描述的权重计算方法来无限扩展,表 3 只列出了每个词的生成树的第一层。这样的生成树构成了该词的上下文向量,并成为词对相似度计算的基础。

4.2 基于 Corpus 的相似度计算实验

本实验考察使用基于训练后的 Corpus 库进行词语相似度计算的效果。

这里选择了几对词语来考察基于 Corpus 的词语相似度计算效果,计算时 κ 取 3(即对每个词 w_i,取其生成树的前 3 层词语来构成该词的上下文特征向量 $d_i = (we_{i1}, we_{i2}, \cdots, we_{in})$),词语的相似度计算结果如表 4。

表 4 词语相似度对比

词语 1	词语 2	相似度	词语 1	词语 2	相似度
篮球	运动	0.21	篮球	宗教	0.01
篮球	教学	0.13	宗教	历史	0.40
篮球	运动员	0.12	宗教	文化	0.24
篮球	电脑	0.01	宗教	电脑	0.06

可以看出,由于词语“篮球”和“运动”的相关关系空间十分相似,因此在两个词的第一层生成树有很多类似的节点,因此这两个词各自构建的上下文特征向量也就较为相似,最终计算得到的相似度较高。反之,“篮球”和“电脑”则相似度较低。

4.3 关系空间裁剪实验

(下转第 644 页)

摘要”。在此基础上,比较了基于词频统计的传统机械摘要方法(系统1)和本文所提出的基于主题划分和句子权重动态调整的摘要方法(系统2)。从表中可以看出,对各类题材的文章,特别是评论类的文档,本系统所得文摘的召回率和准确率都高于传统的摘要方法。尤其是综合指标 F 的值有了显著提升,表明对文档进行的主题划分和文摘句的动态生成方法,很好地兼顾了文摘的覆盖率和准确率,这种优势在文摘压缩率适中(本实验中为10%)时体现得更为明显。尽管划分Web文档的主题消耗了一定时间,但是以划分后的主题为单位抽取文摘时,句子相似度计算的规模大大减小,因而整个系统的运行效率得到了有效提高。

4 结语

鉴于摘要在Web信息检索中的重要作用,本文提出了一种基于网页结构的自动摘要方法。该方法利用网页的层次结构,在主题划分的基础上,充分利用了HTML的标记信息提取主题词,建立向量空间模型,计算句子的重要性,并采用动态调整句子权重的方法生成摘要,克服了传统机械摘要精度不高的缺点。同时按比例分配文摘的方法,解决了多主题文档文摘分布不平衡问题。但是我们也发现,在提取主题词时纯粹以词形为基础,缺乏对不同词语之间语义的理解,造成部分反映文档主题的关键字丢失,从而不能精确地反映文档的主题。

参考文献:

[1] LUHN HP. The automatic creation of literature abstract[J]. IBM Journal of Research and Development, 1958,2(2):159-165.

[2] RUSH JE, SALVADOR R, ZAMORA A. Automatic abstracting and indexing production of indicative abstracts by application of contextual inference and syntactic coherence criteria[J]. Journal of American Society for Information Society, 1971,22(4):260-274.

[3] SALTON G, SINGHAL A, MITRA M. Automatic Text Structuring and Summarization[J]. Information Processing and Management, 1997, 33(2):193-207.

[4] 王永成, 徐慧. OA中文文献自动摘要系统[J]. 情报学报, 1997,16(2):128-132.

[5] RAU LF. Concpetual information extraction and retrieval from natural language input[A]. Proceedings of RIAO 88 Conference[C], 1988. 424-437.

[6] 刘挺,吴岩,王开铸. 基于信息抽取和文本生成的自动文摘系统设计[J]. 情报学报, 1997,16(1):24-29.

[7] DELORT JY, BOUCHON-MEUNIER B, RIFQI M. Enhanced Web Document Summarization Using Hyperlinks[A]. Proceedings of the fourteenth ACM conference on Hypertext and hypermedia[C]. United Kingdom,2003. 208-215.

[8] HU M, LIU B. Mining and Summarizing Customer Reviews[A]. KDD04[C], 2004. 22-25.

[9] 王继成,武港山, 周源远, 等. 一种篇章结构指导的中文Web文档自动摘要方法[J]. 计算机研究和发展, 2003,40(2):398-405.

[10] GUPTA S, KAISER G, NSISTADT D, *et al*. DOM-based Content Extraction of HTML Documents[A]. Proceedings International WWW Conference[C]. New York: ACM Press, 2003. 207-214.

[11] YI L, LIU B, LI X. Eliminating Noisy Information in Web Pages for Data Mining[A]. SIGKDD'03[C], 2003. 24-27.

[12] KIERAS DE. Thematic processes in the comprehesion of technical prose[A]. BRITTON BK, BLACK JB, ed. Understanding Expository Text[C]. Hillsdale, NJ: Lawrence Erlbaum, 1985. 89-107.

[13] KUPIEC J, PEDERSEN J. Atrainable document summarizer[A]. Proceeding of the 18th SIGIR Conference[C],1995. 68-73.

(上接第640页)

为了控制Corpus库的体积,前文提出了关系空间的裁剪,这里分别取容积常数 k 为10,15,20和30对Corpus进行训练,其他常数不变,各项指标取15为基准100,共训练文章6048篇,结果如表5。

表5 关系空间裁剪实验数据对比

容积常数 k	训练时间	Corpus体积
10	40.24	60.46
15	100	100
20	235.41	121.20
50	1413.20	324.24

实验说明,Corpus库的体积随着库存词汇量的增长线性增长,且和容积常数 k 成正比。实验同时发现,当 k 大于15时,继续增大 k,词语相似度的量值会增大,但是对词语亲密程度的排名基本保持不变。

5 结语

本文提出了基于词汇空间和关系空间的Corpus库,用于存储词汇之间的动态关系,并为词语相似度的计算提供数据支持。Corpus库的特点可归结如下:

1) Corpus是对外界语料库中所包含的词语共现信息的结构化整理和优化,可以动态反映词和词之间的共现信息。

2) Corpus的空间开销是受控的,我们使用容积常数 k 来限制节点的平均关系数,使得关系空间的增长和词汇空间的增长成线性关系。

3) Corpus库为词语上下文特征向量的构建提供了数据支持,并可以通过对特征向量的计算来得到词对的相似度量值。

实验证明,Corpus库使用较小的空间存储了较多的有效信息,并为词语相似度的计算提供了足够的数据。

参考文献:

[1] 刘群. 基于<知网>的词汇语义相似度计算[A]. 第三届汉语词汇语义学研讨会[C],2002.

[2] 颜伟,荀恩东. 基于WordNet的英语词语相似度计算[A]. 第二届全国学生计算语言学研讨会[C],2004.

[3] 盛秋艳, 何文广. 基于Hopfield神经网络的概念检索技术[J]. 情报科学,2004,22(3):348.

[4] RIJSBERGEN CJ. Information Retrieval[M]. 2nd edition. London: Butterworths Publishers, 1979.

[5] 罗威. 基于向量空间的中文概念检索技术研究[J]. 情报理论与实践,2003,26(3):226-229.

[6] 胡俊峰. 唐宋诗中词汇语义相似度的统计分析及应用[J]. 中文信息学报,2002, 16(4):39.