语义分析与TF-IDF方法相结合的新闻推荐技术

周 由 戴牡红

(湖南大学软件学院 长沙 410082)

摘 要 在新闻项目的推荐系统中,通常使用TF-IDF权重技术结合余弦相似性度量方法,然而这种技术没有考虑到文字本身的实际语义,因此,提出了基于内容和语义分析相结合的一种新方法。此方法将同义词集合的逆文档频率及语义相似性相结合,采用WordNet同义词集合做相似性计算。构建用户配置文件进行实验测试,验证了该方法的有效性。实验结果表明,提出的语义方法性能优于TF-IDF方法。

关键词 新闻推荐系统,语义分析,语义相似度,WordNet同义词集合

中图法分类号 TP311.132 **文献标识码** A

News Recommendation Technology Combining Semantic Analysis with TF-IDF Method

ZHOU You DAI Mu-hong

(College of Software, Hunan University, Changsha 410082, China)

Abstract Currently in the news item recommendation system, usually using TF-IDF weighting technology combined with the cosine similarity measure, however, this technique does not take into account the actual semantics of the text itself, therefore, the paper propsed a new method based on the combination of contents and their semantic similarities. This method is a collection of synonyms and inverse document frequency combining semantic similarity using WordNet synset do similar calculations. Building user profiles for laboratory tests to verify the effectiveness of the method. Experimental results show that the proposed method outperforms the TF-IDF method.

Keywords News recommendation system, Semantic analysis, Semantic similarity, WordNet synset

1 引言

随着互联网的快速发展,信息呈爆炸式增长,用户很难全面、准确地查找到需要的信息及服务。我们称网上有价值的具体的信息为新闻。然而并非所有的信息都是新闻。通常情况下,新闻分为几个主要的类别,例如商业、体育、政治、科技等等。然而,网站上的新闻并不是所有的用户都感兴趣。在这种情况下,新闻过滤和推荐技术将人们带出了困境,因此在新闻推荐系统和算法方面产生了大量的研究。

基于内容的新闻推荐方法可以分为两种:统计学分析和语义分析[1]。基于统计学的新闻推荐方法是将新闻当成由独立的词语组成的文本,然后通过词语的词频信息将新闻文本建模为高维稀疏的向量,并利用向量间余弦相似度量等方法计算新闻相似性并作出推荐。基于语义分析的新闻相似度量方法则利用特定领域的知识库来构建词语之间的语义关系,以此考察文本之间的相似性。

基于统计学的计算新闻推荐的缺点表现在:需要大规模语料库的支持,忽略了词语之间存在的语义关系,文本表示模型维数高而且稀疏,处理困难。文献[2]提出的基于语义的概念索引检索方法就是由语义网络表示文档内容。文件被映射到WordNet语义网络上,并从一组术语转换成一组概念。随后用TF-IDF等方法提取概念的权重。文献[3,4]都使用了不同程度的语义分析结合TF-IDF相似性比较,然而这种方法往往忽略了同义词组,导致语义分析结果并不那么精确。

针对以上问题,本文提出一种新的新闻推荐方法。结合语义分析的同义词集合的逆文档频率及语义相似性方法(SS-IDF),该方法将基于统计学与语义分析方法相结合,解决了传统语义分析中忽略同义词组的问题,并且加入了语义消岐过程,通过实验证明该方法性能高于传统的TF-IDF方法。

2 TF-IDF新闻推荐方法

传统的基于内容的推荐方法,有很多权重算法,比如逆文档频率(IDF)加权、概率加权、词频(TF)加权等。

TF-IDF方法是一种经典方法,已被普遍用于新闻推荐,并有许多变种。本文选择使用余弦相似性方法结合传统的TF-IDF技术,因为这些技术是众所周知的,并且已被普遍应用,推荐效果相对良好。TF-IDF技术包含两个部分。第一个元素就是TF(t,d),其中t是被计算的单词,d是包含被计算单词的当前文档。TF(t,d)返回一个数字,表示字t出现在文档d多少次。TF-IDF计算中起着至关重要的第二个元素是$idf(t,d)$,被定义为:

$$idf(t,d)=\log\frac{|D|}{|d\in D:t\in T|} \tag{1}$$

本文受湖南省自然科学基金项目(2011FJ3034)资助。

周 由(1988—),女,硕士生,主要研究方向为数据库,E-mail:zhouyou88-com@163.com;戴牡红(1964—),男,副教授,硕士生导师,主要研究方向为数据库技术、企业信息化、决策支持与知识工程。

式中，$|D|$是被比较文档的总数，$|d\in D:t\in T|$是出现 t 的文档总数。本文用到的 TF-IDF 公式为：

$$tf\text{-}idf(t,d)=tf(t,d)*idf(t,d) \tag{2}$$

在计算当前文档的 TF-IDF 向量之后，计算下一个文档的 TF-IDF 向量，一旦每个文档都有一个 TF-IDF 相关的矢量，就可以计算余弦相似。矢量 A 和 B 的余弦相似性度量可以被描述为：

$$similarity(A,B)=\frac{A\cdot B}{\|A\|\cdot\|B\|} \tag{3}$$

式中，A 是一个包含所有用户配置文件中新闻项目中的单词的 TF-IDF 值的向量。B 是一个包含所有未读新闻项目中单词的 TF-IDF 值的向量。$\|A\|$ 是矢量 A 的大小，并且被定义为：

$$\|x\|=\sqrt{x_1^2+\cdots+x_n^2} \tag{4}$$

式中，x 是向量 A 的一个实例，是向量 A 中第 i 被计算的 TF-IDF 值，n 是向量长度。

使用以上介绍的方法为每一个未读文档进行余弦相似性计算，这意味着 A 和 $\|A\|$ 将保持不变，但是 B 和 $\|B\|$ 代表了未读文件的矢量和大小，计算之后所有的结果都将以降序的方式排序，结果与临界值相似的新闻项目将推荐给用户。

然而这种基于内容的方法，纯粹只是对内容计算而不考虑文字的本身含义。基于此，本文提出下面这种基于语义分析的新闻推荐方法。

3 基于语义分析的新闻推荐方法

如图 1 所示，本文提出的基于语义分析的新闻推荐方法包括 3 个部分，第一部分创建用户配置文件，用户配置文件包含所有用户阅读过的新闻项目，根据用户所读新闻即时更新。在配置文件中，每个新闻都有自己的标识符和统一资源标识符。第二部分为本文的核心部分，也是最重要的一部分，即对未读新闻项目进行语义分析，其中包括文本预处理、语义消歧过程、同义词集合创建以及语义相似性度量计算，下节将进行详细介绍。第 3 节根据第 2 节进行比较的结果设置阈值筛选，将符合条件的新闻进行推荐。

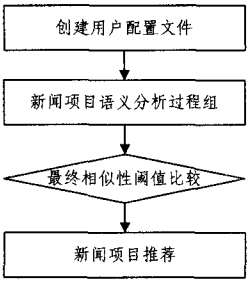

图 1 结合语义新闻推荐处理流程

3.1 语义分析

本文的语义分析研究是基于 WordNet 词汇数据库的。语义分析相似性方法将用户配置文件中词与未读新闻项目中的词进行比较。为了比较这些词的语义相似性，本文使用 WordNet 词典。在 WordNet 中每一个词的形式是一组可能含义的列表，被称为同义词集合。下面进行详细说明。

(1)WordNet 词汇数据库

WordNet 是普林斯顿大学 Miller 等人研制的一部心理语义学词典，包含单词的词义以及各个词义间的关联。传统的词典一般都是按照字母顺序组织词条信息的，它们忽略了词典中同义信息的组织问题。WordNet 正是在这种背景下被提出来的。

WordNet 将词汇分为 5 个大类，即名词、动词、形容词、副词和虚词。实际上，由于虚词通常是作为语言句法的一部分，因此，WordNet 忽略了英语中较小的这部分虚词集。WordNet 最显著的特征就是，它是根据单词的语义而不是词形来组织词汇信息的：每个英文单词对应一个或多个同义词集合，每一个同义词集合表示一个基本的词汇概念。单词和同义词集合之间是多对多的关系，即一个单词对应多个概念，而一个概念又包含了多个单词。本文主要根据 WordNet 词库研究词项的同义词集合做相应相似性计算。

(2)基于 WordNet 的语义分析

基于语义分析的新闻推荐是根据未读新闻项目中词语的 WordNet 同义词集合与用户配置文件的同义词集合进行相似性比较。这些所谓的同义词集合通过使用语义指针根据一个词集到另外一个词集的具体关系进行彼此连接。为了检索一则新闻中一个单词的 WordNet 的同义词集合，首先要明确这个单词的形式、词性标注，以及正确词义[5]。本文使用以下几个步骤来确定 WordNet 单词含义。第一步，文档中所有的单词都需要有一个词性标注。本文使用斯坦福大学的对数线性部分的语音标注器来确定自然语言的词性标注。第二步，删除句子中间断词，根据文献设置好一个间断词列表。第三步，用自适应消岐算法来确定单词的正确含义。因为一个字可以有多种含义，应该选择正确的意思，才能够精确地相互比较不同新闻项目。确定单词含义之后对单词进行 WordNet 同义词集合分配[7]。

用户的配置文件可以表示为一组新闻项目的同义词集合 $S_p=\{s_1,s_2,\cdots,s_m\}$，其中 S_p 是新闻项目 P 的 WordNet 同义词集合，S_i 代表一个 WordNet 的同义词集合新闻，m 是 WordNet 的同义词集合新闻数量。然后将合并用户配置文件中所有新闻项目的 WordNet 同义词集合 R。$R=\bigcup_{p\in P}S_p$。其中 S_p 表示新闻项 P 的同义词集合，P 表示所有以前看过的新闻。未读新闻中同义集的集合被定义为 $U=\{u_1,u_2,\cdots,u_k\}$，其中 u_i 是未读新闻项目的同义词集合，k 是未读消息项的同义词集合的数量。将用户配置文件与未读新闻项目中内容都分配好同义词集合之后，就可以计算相似度进而进行新闻推荐了。

3.2 语义分析相似性计算

语义分析方法要比较用户配置文件中的同义词集合与未读新闻项目的同义词集合的语义相似性。语义相似性计算根据两个集合共有的词性标注进行计算。

首先本文结合 TF-IDF 工作原理，进行相似性计算，式(5)与 TF-IDF 中式(2)不同的是用 s 代替了 t，其中 s 代表一组同义词而不是一个词，其他工作原理不变，利用式(3)计算出未读新闻项目与配置文件相似度。

$$sf\text{-}idf(s,d)=sf(s,d)*idf(s,d) \tag{5}$$

为了进一步计算相似性，一方面要根据未读新闻项目所有可能含义的 WordNet 同义词集合创建 N 维向量，另一方面要与用户配置文件中的同义词集合合并。向量定义为 $V=(\langle u_i,r_j\rangle,\cdots,\langle u_k,r_l\rangle)\ \forall u\in U,r\in R$，其中，$u_i$ 代表未读新闻项目的同义词集合，r_j 代表用户配置文件的同义词集合，k 代

表未读新闻项目的同义词集合的数量，l 代表未用户配置文件的 WordNet 同义词集合的数量。这个向量的一个子集是包含了所有可能词性标注的组合，定义为 $W \in V \forall (u,r) \in W : PSO(u)=POS(r)$。

$POS(u)$和 $POS(r)$分别定义为未读新闻项目中同义词集合和用户配置文件中的同义词集合的可能词性。每个同义词集合表示 WordNet 分类中的一个节点，这种分类是一个层次结构，是一个节点之间的关系。相似性方法旨在说明同义词集合在语义上更接近哪个词集。例如树木在语义上相比于动物词集更接近于植物词集。文献[8,9]进行相似性计算是基于节点的信息内容(IC)。而文献[10,11]进行的相似性计算是基于节点之间的路径长度。

一个节点的信息内容(IC)是同义词集合中所有单词的所有概率总和的负对数。词集中实例 x 的概率定义为 $p(x)$。信息内容可以表示为：

$$IC(s)=-\log\sum_{w\in s}p(w) \tag{6}$$

式中，w 代表同义词集合 s 中的一个单词，它的含义在 s 中给出。

节点路径长度是两个节点之间的最短路径或者是底端节点到顶端节点的最大深度。基于两个节点的路径长度的计算定义为两个节点最近公共子节点。计算公式为：

$$\mathrm{sim}(u,r)=IC(u)+IC(r)-2*IC(LCS(u,r)) \tag{7}$$

基于两个节点路径长度的相似性计算为在两个节点之间的最短路径除以两倍的最大的深度，定义如下：

$$\mathrm{sim}(u,r)=-\log\frac{length(u,r)}{2D} \tag{8}$$

未读新闻的最终相似度是所有组合相似性之和除以组合的总数。最终的相似度定义为：

$$\mathrm{sim}(newsitem)=\frac{\sum_{(u,r)\in W}\mathrm{sim}(u,r)}{|W|} \tag{9}$$

在这里 $\mathrm{sim}(u,r)$是 WordNet 同义词集合 u 和 r 之间的相似性，$|W|$是未读新闻和用户配置文件的同义词集合组合数量。选定阈值后将结合前面相似性计算将两次结果排名都高于阈值的新闻推荐给用户。

3.3 语义相似性算法描述

```
输入:用户配置文件同义词集合向量 U 及未读新闻项目同义词集合
     向量 R
输出:相似度 sim(U,R)
{
  ComB V(u,r);//创建 N 维向量 V
  V=(⟨u_i,r_j⟩,…,⟨u_k,r_l⟩)∀u∈U,r∈R;
  POS(u_i);// 为向量 U 中同义词集的词性生成词性标注
  POS(r_j);// 为向量 R 中同义词集的词性生成词性标注
  Repeat;
  Buid(W);//创建向量 W,W 是 V 的一部分
  For(i;i<=k;i++)
  {
  For(j;j<=l;j++)
  {
  If (POS(u_i)=POS(r_j))
    Add(W)//将所有词性相同的同义词集对加入向量 W 中
  Else
    next();}
    }
    For(i=0;i<=l;i++)
    {
    sim(u_i,r_j);//根据式(8)计算相似性,(u_i,r_j)W
    }
    sim(newsitem);// 最后根据式(9)进行新闻整体相似度比较
}
```

3.4 算法的复杂度

本文利用英文新闻中单词的同义词集信息来反映其语义信息，该算法能够高效、自动地计算出待推荐新闻与用户配置文件在语义层次上的相似关系。算法的复杂度主要体现在式(9)和上述描述算法上，易知其算法复杂度为 $O(n2)$。

4 实验分析

为了验证本文提出的方法的有效性，将和传统的 TF-IDF 方法做比较。

4.1 实验方法

在实验中，利用一个 Web 站点向每位用户显示 100 条新闻项目。新闻全部为英文新闻。用户要根据新闻内容选择是否感兴趣。新闻列表包含了新闻的 URl 地址以及是否感兴趣选项，然后将这些文件作为配置文件进行加载。

实验在多名参与者的配合下进行，他们的年龄都在 20～25 岁之间，所有实验的参与者都是信息学相关领域的学生，以确保他们熟悉预定义配置文件中的相关新闻。在这里要注意的是，由于参与者的数目有限，每个参与者最初都给出了相同的配置，以避免用户偏见。

对于测试结果，采取监督学习的方式，数据被随机分为 60%训练集和 40%测试集。训练集和测试集都包含了一定比例的用户感兴趣的新闻和不感兴趣的新闻。使用训练集时将构造一个用户配置文件，所有感兴趣的新闻项目将被添加，随后计算每个相似性度量值。当相似性值超过设定的临界值时推荐给用户。

为了测量不同相似性算法的性能，需要建立混淆矩阵，用于计算这些方法的性能。进行多次运行，每次随机运行一个测试集和训练集。在实验中，每个用户运行 5 次。收集多个参与者的 5 次测试数据进行分析。

4.2 评测指标

本文使用准确度、精确度、召回率和特异性作为性能指标。使用精确度和召回度计算 F1 平衡指标以及 PR 曲线进行性能比较分析。

4.3 实验结果

处理多名参与者的多次数据，计算出性能衡量指标。以下根据部分实验结果分析性能比较。

表 1 所列为部分实验参与者信息。表 2 列出 TF-IDF 和 SS-IDF 方法在 cutoff＝0.65 时的各项性能指标，如表中所示，SS-IDF 优于 TF-IDF。如图 2 所示，SS-IDF 方法的 F1-度量值明显优于传统的 TF-IDF 方法。

除此结果分析之外，本文还采用了 PR(Precision-Recall)精密召回曲线进行实验结果分析。这条曲线表明一个推荐算法的精度和召回率之间的关系。通常情况下，在一个较低的召回率条件下可以获得较高的精度。根据图 3 曲线所示，SS-IDF 明显优于 TF-IDF，在相同的召回率下，SS-IDF 具有更高的精度。例如，在召回率为 70%的情况下，SS-IDF 精度为

(下转第 300 页)

plex scene rendering[J]. Visual Languages and Computing, 2005,16(5):455-479

[11] Laine S. A general algorithm for output-sensitive visibility preprocessing[C]//Proceedings of the 2005 Symposium on Interactive 3D Graphics and Games. 2005:31-40

[12] Djeu P, Keely S, Hunt W. Accelerating shadow rays using volumetric occluders and modi_ed kd-tree traversal[C]//Proceedings of the Conference on High Performance Graphics 2009. 2009:69-76

[13] Kao C, Tsai R. Properties of a level set algorithm for the visibility problems[J]. Scientific Computing, 2008, 35(2):170-191

[14] Choi B, Chang B, Ihm I. Technical section: Construction of efficient kd-trees for static scenes using voxel-visibility heuristic [J]. Computers and Graphics, 2012, 36(1):38-48

[15] Jackins C, Tanimoto S L. Oct-trees and their use in representing 3-d objects[J]. Computer Graphics and Image Processing, 1980, 14:249-270

[16] Reinhard E, Smits B, Hansen C. Dynamic acceleration structures for interactive ray tracing[C]//Eurographics Workshop on Rendering. 2000:299-306

[17] Luebke D, Erikson C. View-dependent simplification of arbitrary polygonal environments // SIGGRAPH 97 Proc. August 1997: 199-208

[18] Xia J C, Varshney A. Dynamic view-dependent simplification for polygonal models[C]//Proceedings of Visualization '96. October 1996:327-334

[19] Garland M, Willmott A, Heckbert. Hierarchical face clustering on polygonal surfaces[C]//Proceedings of the ACM Symposium on Interactive 3D Graphics. 2001:49-58

[20] Barber C B, Dobkin D P, Huhdanpaa H T. The quickhull algorithm for convex hulls[J]. ACM Trans. on Mathematical Software, 1996, 22(4):469-483

[21] Wong S K, Cheng Yu-chun, Lii S Y. GPU Ray Tracing Based on Reduced Bounding Volume Hierarchies[C]//Proceedings of the Ninth International Conference on Computer Graphics, Imaging and Visualization. 2012:1-6

[22] Wang Rui, Qian Xue-lei. OpenSceneGraph 3 Cookbook[M]. Packt Publishing, March 2012

(上接第 269 页)

82%，而 TF-IDF 仅仅为 68%。因此我们可以得出结论，SS-IDF 比 TF-IDF 在精度与召回率方面具有更好的平衡性。

表 1 部分实验参与者信息

用户	性别	年龄	专业	研究方向
User1	男	25	计算机科学与技术	J2EE 架构及技术
User2	女	22	软件工程	大数据处理
User3	男	23	信息与通信工程	无线通信技术
…	…	…	…	…
User8	男	24	计算机技术	人工智能 …
…	…	…	…	

表 2 TF-IDF 算法和 SS-IDF 算法在 cutoff=0.65 平均实验结果

性能指标	IF-IDF	SS-IDF
Accuracy	78.3%	80.8%
Precision	77.2%	77.9%
Recall	21.8%	35.7%
Specificity	97.5%	94.6%
F1-measure	38.9%	48.6%

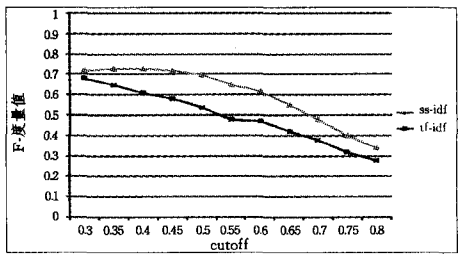

图 2 TF-IDF 算法和 SS-IDF 算法在不同 cutoff 下的 F-性能指标

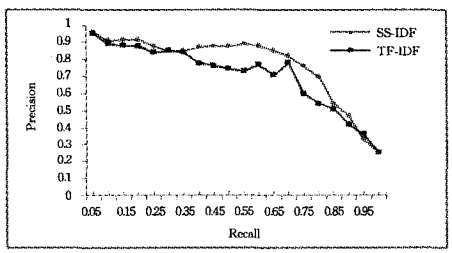

图 3 PR 曲线

结束语 为了改善传统上用于新闻项目推荐的词频逆文档频率(TF-IDF)加权技术，本文提出了一种新方法。本文的方法基于概念和它们的语义相似性，从中得到新闻项目之间的相似性。在新闻项目推荐中通过比较性能指标，本文提出的算法性能优于传统的词频逆文档频率(TF-IDF)加权技术。

参 考 文 献

[1] 华秀丽，朱巧明，李培峰. 语义分析与词频统计相结合的中文文本相似度量方法研究[J]. 计算机应用研究，2011，29(3)：834-836

[2] Goossen F, Jntema W, Frasincar F, et al. News Personalization using the CF-IDF Semantic Recommender[C]//Proc of the International Conference on Web Intelligence, Mining and Semantics. 2011

[3] 黄承慧，印鉴，侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似性度量方法[J]. 计算机学报，2011，34(5)：857-863

[4] 李明涛，罗军勇，尹美娟，等. 结合词义的文本特征权重计算方法[J]. 计算机应用，2012，32(5)：1355-1358

[5] Toutanova K, Klein D, Manning C D, et al. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network[C]//Proc of " NAACL". 2003:173-180

[6] Jensen A S, Boss N S. Dty similarity[OL]. http://damn. dk/similarity/javadoc/model/similarity/Lesk. html, 2008

[7] Lextek: Onix Text Retrieval Toolkit {API Reference. http://www. lextek. com/manuals/onix/stopwords1. html (2011)(stop word)

[8] Jiang J J, Conrath D W. Semantic Similarity Basedon Corpus Statistics and Lexical Taxonomy[J]. Proc of 10th International Conference on Research in Computational Linguistics, 1997, 19 (33)

[9] Fellbaum C. WordNet: an electronic lexical database [OL]. WordNet is available from http://www. cogsci. princeton. edu/wn, 2010

[10] Resnik P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy[C]//Proc of the 14th International Joint Conference on Artificial Intelligence. 1995, 11:448-453

[11] Wu Zhi-biao, Palmer M. Verb Semantics and Lexical Selection [C]//Proc of 32nd Annual Meeting on Association for Computational Linguistics. 1994:133-138