基于模糊集和 RSS 的 Web 教育资源 Rank 算法

王 杨,杨娜娜,陈付龙,赵传信
(安徽师范大学 数学计算机科学学院,安徽 芜湖 241000)

摘 要:随着 Web 教育资源指数级增长以及受污染程度的加剧,如何汇聚异构网络环境下面向用户的 Web 教育资源成为新的挑战。为了使终端用户能够获得高效有序的 Web 教育资源,文中提出了一种基于模糊集和 RSS 的 Web 教育资源 Rank 算法,并进行了相关分析。算法首先通过模糊集中的 Euclid 模糊度刻画查询内容与资源之间的模糊关联度;其次采用 RSS 聚合技术快速汇聚用户需要的 Web 教育资源;最后基于中国知网数据集的实验表明,该算法能满足 Web 教育资源终端用户个性化资源获取的需要。

关键词:模糊集;RSS;Web 教育资源;Rank 算法
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2013)02-0127-04
doi:10.3969/j.jssn.1673-2013.02.032

Rank Algorithm of Web Educational Resources Based on Fuzzy Sets and RSS

WANG Yang, YANG Na-na, CHEN Fu-long, ZHAO Chuan-xin
(School of Mathematics and Computer Science, Anhui Normal University, Wuhu 241000, China)

Abstract: With the exponential growth of Web educational resources and the degree of contamination increased, how to aggregate user-oriented Web educational resources has become the new challenge in the heterogeneous network environment. In this paper, present a kind of Web educational resources Rank algorithm based on fuzzy sets and RSS in order to make terminal users obtain efficient Web educational resources ordered. First of all, the algorithm depicts the fuzzy association degree between the query and the resources through the Euclid ambiguity of fuzzy sets. Second, the use of RSS aggregation technologies rapidly aggregates the Web educational resources users needs. Finally, the experiment of CNKI data set shows that the algorithm can meet the needs of personalized educational resources access.

Key words: fuzzy set; RSS; Web educational resource; Rank algorithm

0 引 言

Web 教育资源是指以数字信号在互联网上共享的具有教育价值的各类信息资源[1,2]。形式多样的 Web 教育资源主要分布在:

(1)大型的搜索网站(如 Google、Baidu 等);

(2)付费的数字图书馆(如中国知网、万方数据库等);

(3)教育服务提供商(如新东方、英孚教育等)。

面对日益丰富的教育资源,用户不仅需要透明的协同与共享,又需要有效地选择和汇聚“以用户需求为中心”高可信的教育资源手段。

异构网络 Web 教育资源汇聚和选择是指以开发利用网络信息资源为基础,根据教育主体的需求,运用电子、信息技术对不同自治域中的相关知识内容进行采集、加工、存储、传输、检索和利用,以新的序列化的知识单元呈现给教育终端用户的服务性工作。根据用户群体,现有的 Web 资源排序可分为三大类。

一是面向社会群体用户的排序算法,如 PageRank 算法[3~5]、HITS 算法[6,7]等;

二是面向组用户的排序算法,如 Direct Hit 算法[8]、基于群体特性的排序算法[9]等;

三是面向个性化用户的排序算法,如 BM25 加权算法[10]、个性化排序算法[11]等。

总体来说,上述方法均难以满足用户获取领域相关、时间相关、内容优质的高可信 Web 教育资源的需求。因此,文中提出了基于模糊集和 RSS 的 Web 教育资源 Rank 算法。

1 相关理论与技术

鉴于模糊集中的 Euclid 模糊度可以刻画查询内容

收稿日期:2012-06-03;**修回日期**:2012-09-05
基金项目:2011 年教育部人文社会科青年基金项目(11YJC880119)
作者简介:王 杨(1971-),男,教授,博士后,主要研究方向为可信计算、教育技术。

与资源之间的模糊关联度,RSS 文档聚合技术可以快速汇聚用户需要的教育资源,因此,采用其作为 Web 教育资源 Rank 方法的理论基础和技术手段。

1.1 模糊集

给定一个论域 U,从 U 到 $[0,1]$ 的一个映射 $\mu_A: U \rightarrow [0,1]$,设 $A = \{\mu_A(u) \mid u \in U\}$,则称 A 为论域 U 上的一个模糊集,函数 μ_A 为模糊集 A 在论域 U 上的一个隶属函数,$\mu_A(u)$ 为 u 对模糊集 A 的隶属度[12]。

鉴于模糊数学的原理,隶属函数可以对资源进行评估。根据模糊集中的隶属函数,可以刻画查询内容与教育资源之间的隶属度。隶属函数定义如下:

定义 1:隶属函数定义为

$$\mu_A(u_i) = \frac{N(u_i)}{\sum_{0 \leqslant j \leqslant n} N(u_j)} \tag{1}$$

其中 u_i 为代表教育资源的词集或要查询的内容,n 为 u_i 的个数,$N(u_i)$ 为 u_i 在教育资源中的数量,而 $\mu_A(u_i)$ 的值为 u_i 对模糊集 A 的隶属度。

Euclid 模糊度通过隶属度来确定查询内容可以模糊代表教育资源的模糊关联度。Euclid 模糊度定义如下:

定义 2:Euclid 模糊度定义为

$$d(A) = \frac{2}{\sqrt{n}} \sqrt{\sum_{i=1}^{n} \mid \mu_A(u_i) - \mu_{A_{0.5}}(u_i) \mid^2} \tag{2}$$

其中,

$$\mu_{A_{0.5}}(u_i) = \begin{cases} 1 & \mu_A(u_i) \geqslant 0.5 \\ 0 & \mu_A(u_i) < 0.5 \end{cases} \tag{3}$$

1.2 RSS 技术

简易聚合(Really Simple Syndication,RSS)是根据用户的需要,把相关站点的教育资源高效聚合,把预定信息(标题、摘要、内容)“推送”到用户桌面[13]。

由于 RSS 技术是根据用户的需要,则需要自定义用户标签。RSS 标签定义如下:

定义 3:RSS 标签定义为

$$Tag = \{Tag_1, Tag_2, Tag_3, \cdots, Tag_m\} \tag{4}$$

其中,Tag_i 为刻画 Web 教育资源的第 i 个关键属性,$C[Tag_i]$ 为用户自定义的对 Tag_i 的关注度,$(0 \leqslant C[Tag_i] \leqslant 1)$。

RSS 技术需要根据用户定义的标签,对教育资源集计算其资源标签比重。资源标签比重定义如下:

定义 4:资源标签比重。设教育资源集为 $ER = \{er_1, er_2, \cdots, er_n\}$,资源 er_i 中 Tag_j 的数量为 $N_{er_i}[Tag_j]$,则 RSS 标签 Tag_j 在资源 er_i 中所占的比重,即

$$C_{er_i}[Tag_j] = \frac{N_{er_i}[Tag_j]}{M_{er_i}} \tag{5}$$

其中,

$$M_{er_i} = \sum_{1 \leqslant j \leqslant 10} N_{er_i}[Tag_j] \tag{6}$$

又根据用户自定义的标签比重,计算出用户最满意的教育资源,再推送给用户。

定义 5:用户满意度。对于资源 er_i,用户满意度定义为

$$W_{er_i} = \sum_{1 \leqslant j \leqslant 10} \frac{C_{er_i}[Tag_j]}{C[Tag_j]} \tag{7}$$

其中,W_{er_i} 的大小体现用户对 Web 教育资源满意程度的高低。

2 基于模糊集和 RSS 的 Web 教育资源 Rank 方法

教育资源检索涉及三方面的内容:教育资源集的表示、教育资源终端用户查询的表示、用户查询与教育资源集的匹配及其排序。文中教育终端用户除了输入标签和查询的内容外,还需要输入标签和查询内容对用户所需要的重要程度,重要程度分为十个等级,等级越高,表示越重要。

2.1 算法描述

算法 1:基于模糊集的 Rank 算法。

Input: $Q = \{q_1, q_2, \cdots, q_s\}$;//用户查询内容

$W_Q = \{w_{q1}, w_{q2}, \cdots, w_{qs}\}$;//用户查询内容的比重

Output: γ_f ;//通过模糊集理论查找后的教育资源排序结果

BEGIN

for er_1 to er_n

{

extract $U = \{u_1, u_2, \cdots, u_t\}$;

for u_1 to u_t

$\mu_A(u_i) \leftarrow N(u_i) / \sum_{0 \leqslant j \leqslant n} N(u_j)$;//关键词的隶属度

$d(er_j) \leftarrow \frac{2}{\sqrt{s}} \sqrt{\sum_{i=1}^{s} \mid \mu_A(q_i) \times w_{q_i} - \mu_{A_{0.5}}(q_i) \mid^2}$;

//用户查找内容可以模糊代表教育资源 er_j 的模糊度

$\gamma_f \leftarrow \alpha d(er_j)$

return γ_f ;

}

END

算法 2:基于 RSS 技术的 Rank 算法。

Input: $Tag = \{Tag_1, Tag_2, \cdots, Tag_m\}$;//用户定义标签

$W_{Tag} = \{w_{Tag_1}, w_{Tag_2}, \cdots, w_{Tag_m}\}$;//用户标签比重

Output: γ_r ;//通过 RSS 技术推送给用户的教育资

源排序结果

BEGIN

for er_1 to er_n

for Tag_1 to Tag_m

{

$N_{er_i}[Tag_j] \leftarrow \text{count}[Tag_j]$;//计算教育资源 er_i 中 Tag_j 的数量

$C_{er_i}[Tag_j] \leftarrow N_{er_i}[Tag_j] / M_{er_i}$;//计算教育资源 er_i 中 Tag_j 的标签比重

for j: 1 to m

$W_{er_i} \leftarrow (C_{er_i}[Tag_j] / C[Tag_j]) + W_{er_i}$;//计算教育资源 er_i 的用户满意度

$\gamma_r \leftarrow (1 - \alpha) W_{er_i}$

return γ_r ;

}

END

2.2　算法分析

上述算法的主要步骤包括以下两点。首先,从教育资源集 ER 中提取能够代表这个资源的词集 $U = \{u_1, u_2, \cdots, u_t\}$ 。根据隶属函数算出每个检索词的隶属度。根据查询的内容及其比重,利用 Euclid 模糊度计算查询内容可以模糊代表此教育资源的模糊度。然后对其检索出的教育资源根据目前用户所定义的 RSS 标签及其标签比重计算标签与教育资源之间的关联度。综合比较,对检索出的文档进行排序。基于模糊集和 RSS 的 Web 教育资源 Rank 算法主要分为两个算法:算法 1 是通过模糊集中的隶属度函数和 Euclid 模糊度来实现资源的排序,时间复杂度是 $O(n^2)$ 。算法 2 是根据 RSS 技术原理推送出用户满意的资源,时间复杂度是 $O(n)$ 。

3　实验结果与分析

实验主要基于中国知网数据集并与中国知网的排序结果从资源标签比重、用户满意度、模糊度等方面进行比较分析。

首先,通过在中国知网中查询“模糊集”,共有 2937 条记录。然后随机抽取若干条记录作为实验测试数据集。随机抽取的数据集部分如表 1 所示。

实验 1:模糊度的影响分析实验。

用户给出查询的内容为“模糊集”、“隶属函数”,其查询内容的比重分别为 0.8、0.6。根据算法 1 算出查询内容可以代表教育资源的模糊关联度。

图 1 中的条状图代表用户查询的内容可以模糊代表资源的关联度,条状图越高说明该资源与用户要查询的教育资源越接近,越低代表该资源与查询内容之间的关联越小,零点代表此教育资源与查询内容无关。

表 1　实验数据集及其在 CNKI 排序结果

编号	教育资源集	知网排序结果
er1	关于模糊集的粗糙度	28
er2	一种覆盖粗糙模糊集模型	33
er3	基于模糊集隶属度特征和贴近度的徽标识别	34
er4	直觉模糊集的包含度	35
er5	模糊集间的语义关联度及其应用	52
er6	模糊集与粗糙集	74
er7	基于模糊集的信息检索方法	104
er8	基于模糊集的术语搜索方法研究	191
er9	直觉模糊集上的混合单调包含度	221
er10	模糊集的包含度与熵	293
er11	基于模糊集理论的关联规则研究	368
er12	基于相似 Rough 集的模糊检索策略	922
er13	基于隶属度函数的决策层融合算法	1164
er14	基于隶属度的数据库模糊结果排序方法	1203

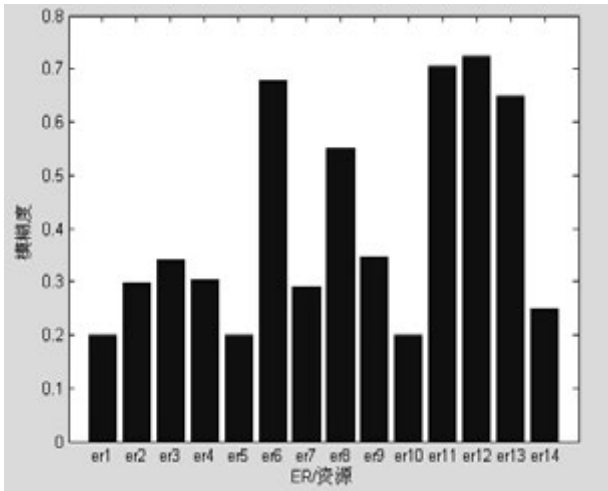

图 1　用户查询的 Web 教育资源内容的模糊度

实验 2:标签比重比较实验。

由于教育资源终端用户在某一段时间内只针对于某一特定领域获取教育资源,因此用户自定义标签并给出用户标签比重。根据算法 2 计算出实验数据的资源标签比重,聚集出用户满意的教育资源。假设用户自定义 RSS 标签为{教育资源,模糊集,RSS,排序,标签,隶属函数,包含度,0,0,0},其对应比重为{1,0.9,0.9,0.6,0.4,0.6,0.7,0,0,0 }。根据公式(5)计算出实验数据的资源标签比重如图 2 所示:

图 2 中表明各教育资源的标签比重与用户自定义标签比重相比较,各个标签的比重与用户自定义标签比重距离越接近,资源用户越满意。

实验 3:用户满意度。

根据模糊集中查询内容模糊代表教育资源的模糊度和模糊满意度进行综合比较,由公式(7)计算出实验数据的用户满意度如图 3 所示。图 3 中的数据集排序结果表明用户标签比重与用户满意度之间关联性并不强,需要通过一定的方式进行调节。

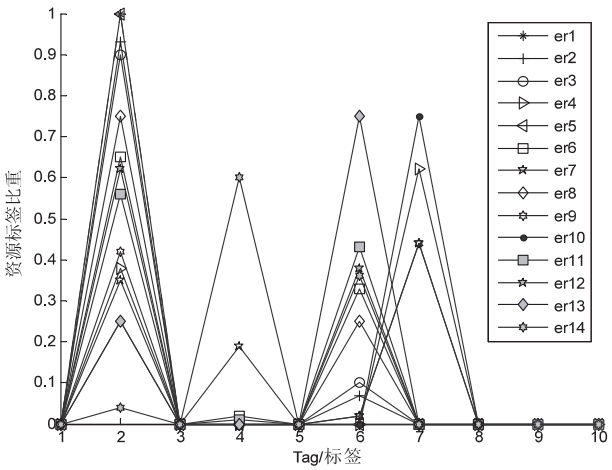

图 2 资源标签比重计算结果

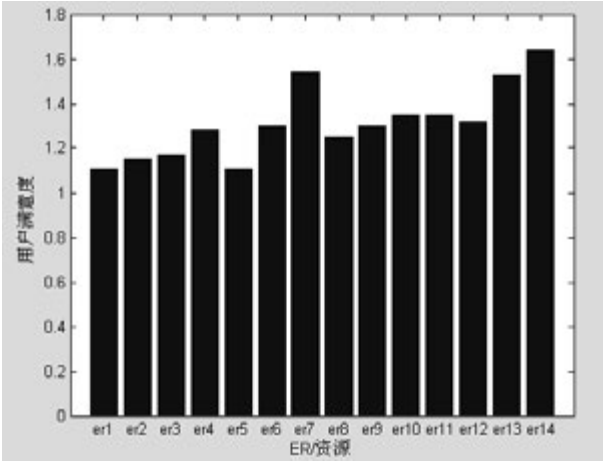

图 3 用户满意度

图 4 表明 er13,er11,er12,er6,er14,er7,er8,er9,er4,er10,er3,er2,er1,er5 是根据文中算法的相应排序结果。根据图 4,可以看出中国知网中的排序结果与用户需要的教育资源排名结果之间存在不一致性,有一些用户需要的一些教育资源在中国知网中排名靠后,但是经过文中的算法,可以把用户需要的资源排名提前。所以文中算法在一定程度上更符合用户的个性化需求。

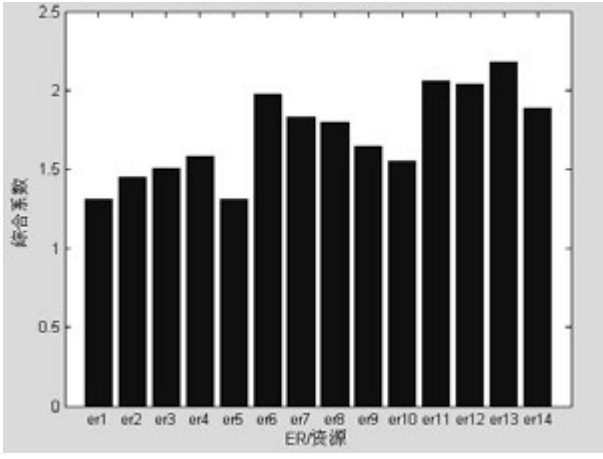

图 4 综合排名结果

4 结束语

文中主要提出了一种基于模糊集和 RSS 的 Web 教育资源 Rank 算法。算法使用 Euclid 模糊度作为衡量查询内容可以模糊代表教育资源的程度,并通过 RSS 技术聚合教育资源终端用户自己相关领域的,高可信的教育资源。基于中国知网数据集的实验表明,文中算法能够把教育资源终端用户所需要的教育资源排序提前,为用户推送“用户所需要的”优质资源。下一步工作中,将针对基于中心资源和中心用户影响度的 Web 教育资源聚合问题展开研究。

参考文献:

[1] Akhshabi M, Khalatbari J. Educational Standard Content Design System for Virtual University[J]. Procedia-Social and Behavioral Sciences, 2011, 28: 855-861.

[2] Mustakerov I, Borissova D. A conceptual approach for development of educational Web-based e-testing system[J]. Expert Systems with Applications, 2011, 38(11): 14060-14064.

[3] Wu Gang, Wei Yimin. An Arnoldi-extrapolation algorithm for computing PageRank[J]. Journal of Computational and Applied Mathematics, 2010, 234(11): 3196-3212.

[4] Yan Lili, Gui Zhanji, Du Wencai. An Improved PageRank Method Based on Genetic Algorithm for Web Search[J]. Procedia Engineering, 2011, 15: 2983-2987.

[5] Boldi P, Santni M, Vigna S. PageRank as a function of the damping factor[C]//Proceedings of the 14th International Conference on World Wide Web. USA: ACM Press, 2005: 557-566.

[6] Klenberg J M. Authoritative sources in a hyper-linked environment[J]. Journal of the ACM, 1999, 46(5): 604-632.

[7] Bharat K, Henzngerm R. Improved algorithms for topic distillation in a hyper-linked environment[C]//The 21st annual international ACM SIGIR conference. USA: ACM Press, 2005: 186-193.

[8] The Dirct Hit Popularity Engine Technology[S]. 1999.

[9] Hersovici M, Jacovi M, Maarek Y S. The shark-search algorithm-An application[C]//Proceedings of the 7th International World Wide Web Conference. Brisbane, Australia: ACM Press, 1998: 317-326.

[10] Zhai Chengxiang. Statistical language models for information retrieval a critical review[J]. Foundations and Trends in Information Retrieval, 2008, 2(3): 137-213.

[11] Kang Fuwei, Liu Xiaodong. A Personalized Ranking Approach via Incorporating Users' Click Link Information into PageRank Algorithm[J]. Energy Procedia, 2011, 13: 275-284.

[12] 刘普寅,吴孟达. 模糊理论及其应用[M]. 长沙:国防科技大学出版社,1998.

[13] 刘清堂,白新国. 基于 RSS 的教育资源服务系统研究[J]. 计算机工程与设计,2008,29(2):474-476.

基于模糊集和RSS的Web教育资源Rank算法

作者： 王杨， 杨娜娜， 陈付龙， 赵传信
作者单位： 安徽师范大学 数学计算机科学学院,安徽 芜湖 241000
刊名： 计算机技术与发展
英文刊名： Computer Technology and Development
年，卷(期)： 2013(2)