

一种基于舆情敏感新信息的发现和搜索方法

王 文^{1,2}, 王树锋^{1,2}, 朱锡芳¹

(1 常州工学院, 江苏 常州 213002; 2 常州市软件技术与应用重点实验室, 江苏 常州 213002)

摘 要: 提出了一种基于舆情的敏感新信息搜索方法, 通过对一定时间、区间内查询文档增长数据的监控, 判断哪些信息属于敏感新信息, 并利用新信息的多个时间特征, 结合传统搜索引擎下根据检索条件得到的检索结果, 实现包含新信息数据的敏感新信息排序模型, 解决了新信息数据不足的问题. 实验表明, 该模型既保证常规信息的相关性又很好地实现了对新信息排序.

关键词: 敏感新信息; 敏感新信息搜索; 新信息时间特征; 新信息排序模型

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2013)12-0113-04

A Method of Discovering and Searching Sensitive Recency Information Based on Public Opinion

WANG Wen^{1,2}, WANG Shu-feng^{1,2}, ZHU Xi-fang¹

(1 Changzhou Institute Technology, Changzhou 213002, China;

2 Changzhou Key Laboratory of Software Technology and Application, Changzhou 213002, China)

Abstract: The recency sensitive information search refers to the procedure of ranking Web document by considering the information time as the correlation index, sorting new information in the large-scale Web search is a very difficult problem. A method of search sensitive recency information is proposed. By monitoring the query document growth over a period of time, the sensitive information is determined. Combined multiple time feature of sensitive information with the retrieval results obtained by traditional search engine, the new sort model using new information data is achieve and solves the problem of insufficient information data. Experiments show that the model ensure the correlation routine information and achieve a good sort of new information.

Key words: recency sensitive information; recency sensitive information searching; recency information temporal feature; recency sorting model

1 引言

舆情是指在一定的社会空间内, 围绕中介性社会事件的发生、发展和变化, 民众对社会管理者产生和持有的社会政治态度. 对于网络舆情监管而言, 首要解决的问题是如何在 Web 页面中迅速判断突发舆情信息(称为敏感信息)并持续获取最新的舆情信息(称为新信息).

敏感新信息排序是指将信息出现的时间因素作为相关性指标对 Web 搜索得到的文档进行的排序,

在大规模 Web 搜索中新信息排序是一个非常困难的问题. 敏感新信息排序算法的相关性评价指标要包含标题相关性和信息的时间因素, 为了保证检索系统检索性能不降低, 系统只有在需要的时候才增加对新信息的检索; 其次, 文档的新鲜度对时间敏感, 查询爆炸式新闻要将一天或几小时内的相关性结果显示出来, 而新信息查询操作是不同的. 再次, 时间特征通常很难从 Web 页面中抽取出来, 度量文档的生存时间很难. 最后, 训练数据的收集和评价需要与检索组在时间上同步, 并且要体现出对时间上

收稿日期: 2013-01-20; 修回日期: 2013-03-10

基金项目: 江苏省教育厅自然科学基金基础研究基金(1KJD520001); 常州市应用基础计划研究项目(CJ20120009)

的敏感.

文献[1]中研究了 Web 动力学、Web 结构和页面质量的关系,证明 PageRank 算法对新页面是有偏见的.文献[2]中提出的 T-Rank Light 和 T-Rank 算法,通过分析页面的链接,确定页面及其链接页面升级和最近新信息升级的时间戳的关系.文献[3]中除了强调页面的动力学特性是排序的重要因素外,也考虑了用马尔可夫链下静态概率分布表示的页面新鲜度,提出了 TS-Rank 算法.文献[4]中提出利用时间标识的方法,通过聚集匹配信息和其中包含的时间标识信息,对与时间有关问题回答的结果进行了改进.

文献[5-6]中提出在 Web 搜索中改进排序新信息的方法.

新信息查询分类在新信息排序中承担重要的角色,文献[6]中利用点击显示查询结果的预测概率确定查询得到的新闻的价值.文献[7]中利用满足适应突发新闻事件需求的监督模型,对特定新闻搜索结果的点击率进行估计.

2 敏感新信息的发现方法

当某个主题成为敏感新信息(如突发新闻),对该主题的查询就表示出异常,在查询流中该主题出现的次数会越来越多,同时在新闻文章中该主题出现的次数也越来越多.本文提出的敏感新信息发现方法的基本思想是:根据一个时间段内用户提交的查询数据和该时间段内发布的新闻,采用 n 元语言模型分别进行建模.具体讲就是针对每一个时间段收集搜索引擎在该时间段接收到的查询数据,同时收集该时间段内发布的新闻文章.通过估计两个 n 元语言模型 $M_{C,t}$ 和 $M_{Q,t}$,分别表示时间 t 范围内的内容数据和查询数据.只需要比较 t 时刻的相关模型与 t 时刻之前的相关模型,就能够确定在时间 t 某个查询 q 是不是敏感新信息查询.

该方法执行的过程为:

第一步,以 r_i 为当前模型的起点,选择几个固定的时间区间,在这些区间内定义几个参考模型,表示在 $t-r_i$ 时间段内的内容和查询数据,这些模型表示为 $M_{C,t-r_i}$ 和 $M_{Q,t-r_i}$.

第二步,根据当前查询模型,计算 q 产生的可能性 $P(q | M_{Q,t})$.该模型采用线性插值平滑低维的 n 元语言模型.用同样的方法计算,得到参考模型 $P(q | M_{Q,t-r_i})$ 的多个值.

第三步,根据查询流数据建立的模型,计算查询

的数量:

$$\text{buzz}(q, t, Q) = \max_i \log P(q | M_{Q,t}) - \log P(q | M_{Q,t-r_i}).$$

根据同样的过程计算文章内容,得到一个类似的内容模型: $\text{buzz}(q, t, C)$.组合这两个值,得到最终的值为

$$\text{buzz}(q, t) = \lambda \cdot \text{buzz}(q, t, Q) + (1 - \lambda) \cdot \text{buzz}(q, t, C).$$

如果最终值超过某个阈值 k ,则认为该查询就是敏感新信息查询,查询的信息就是敏感新信息.

3 新信息的时间特征

Web 页面的新旧是该页面存在的时间,表示为: $\Delta t = t_{\text{query}} - t_{\text{page}}$,其中, t_{query} 是用户查询的时间, t_{page} 是页面时间,页面时间包括页面创建的时间和页面内容中包含的时间信息.通常页面时间很难轻易获得.对于一个 Web 页面,通过页面内容中的时间信息和页面的链接关系中的时间信息,可以确定页面时间,页面发现时间和页面创建时间通常是不相同.

页面中的与时间有关的特征可以决定该页面是不是包含新信息.这些特征有:时间戳特征、连接时间特征、页面更新数量特征和页面分类特征.

时间戳特征是从页面内容中抽取以某种格式出现的时间值,利用常用的时间表示格式,从给定的文档中找到对应的时间,对这些时间数据进行分析,计算时间戳数量、最小时间戳(对于最老的页面)、最大时间戳(对于最新的页面)、时间戳中的时间平均值和标准偏差等数据,推算页面存在的时间.

连接特征从基于连接的页面发现注册信息中保存的页面存在的时间.当 Web 上发现一个连接,系统记录连入该 URL 的时间.跟时间戳特征一样,计算页面 URL 的链接时间,包括数量、最小值、最大值、平均值和标准偏差,来推断出该页面的时间.由于网络爬虫算法和效率的影响,一个页面从创建到发现会花费很长的时间.虽然连接时间不一定能准确表示页面的存在时间,但可以提供强烈的页面存在时间的证据.

页面更新数量特征是利用网络爬虫程序得到的 URL 链接结构、多样性和新信息等计算得到的,可以描述 URL 页面的流行程度.

时间戳特征和连接时间特征用来表示页面的新鲜程度.页面更新数量特征是页面和连接的更新频率,反映了新页面的流行程度.页面分类特征也与新

页面有关. 与新信息相关的单个特征不能直接和准确表示页面的新鲜程度,但是这些特性提供了有用的证据. 在排序模型学习阶段,将不同的新信息特征组合起来,可以很好地表示页面的新鲜程度.

4 新信息排序模型

构建模型就是根据某种度量指标,从偏好数据集中利用机器学习技术学习得到一个最优的排序函数,使得信息损失最少. 在新信息排序中,可以利用的信息有:新信息排序数据集、新信息特征数据、常规排序数据集以及常规排序数据特征. 这些数据是有差别的,新信息排序数据仅仅用于对新信息进行查询,而常规排序数据能适应任意的查询. 常规排序数据不一定具有新信息的特征值. 两种排序数据的评价不一样,新信息数据对时间更敏感. 新信息排序数据包含(查询条件、url、相关度数值)等信息,每个url表示为一组特征向量,包括基于连接结构的特征,基于内容的特征和基于点击的特征等. 新信息排序数据收集和评价十分困难,这会严重影响排序模型的学习和构建. 虽然常规排序数据集与新信息数据集有差异,但这也是一个非常重要的构建新型模型的数据资源,可以用来帮助学习新信息排序模型.

最理想的构建新信息排序模型的方法是用新信息排序数据训练该模型,特征集是常规排序特征和新信息特征的组合. 我们把这种模型作为评价其他模型的基础,这种称为基础模型.

第一种模型称为复合排序模型,将常规排序模型得到的相关性排序值作为一个特征,添加到新信息特征中,用新信息排序数据作为训练数据,训练学习新信息模型. 这种模型利用了包含查询URL对的相关性信息的常规排序模型的信息,同时特征维度大大减少,降低了产生过学习的可能性. 因为训练数据相对小,特征维度也较低,该模型的训练速度很快.

第二种模型称为权重模型,将新信息排序数据和常规排序数据组合起来作为训练模型的数据,两个数据集以不同的比例进行组合. 特征集是两个特征集的组合. 由于数据充分,该模型不会过学习. 本文将新信息数据和常规数据按照1:1、1:2和1:3的比例组合,构建新的训练数据,训练模型.

第三种模型成为迁移学习模型,利用常规相关性训练数据构建基本模型,该基本模型根据新信息排序数据进行修正.

GBrank算法采用决策树梯度集成技术,基本

思想是,对于偏好集 $S = \{ \langle x_i, y_i \rangle \mid x_i \succ y_i, i = 1, 2, \dots, N \}$ 中的偏好对 $\langle x, y \rangle$, x 的排序比 y 的排序高. GBrank算法中的学习排序函数计算排序函数 h 使偏好集中尽可能多的数据满足下列的条件: 如果 $x \succ y, i = 1, 2, \dots, n$, 则 $h(x_i) \geq h(y_i)$. 用下面损失函数度量给定的排序函数 h 的风险.

$$R(h) = \frac{1}{2} \sum_{i=1}^N (\max \{0, h(y_i) - h(x_i) + \tau\}^2)$$

以上模型用 GBrank 算法^[12] 作为基本模型,利用新信息偏好数据调整该模型的参数,使之尽可能优化排序顺序,目标是尽可能减少损失函数.

5 新信息排序模型评价

数据集:在四个月的时间内收集了 70 131 查询URL对新信息数据. 对于每一个查询URL对,在同一天进行评价和特征抽取. 将数据集随机地划分为训练数据和测试数据,其中训练数据为 41 678 对,测试数据为 28 453 对.

由于本文提出的模型需要常规排序数据的支持,同期收集了 282 927 查询URL对,这些数据的新信息特征不一定准确,因为数据集中的数据值不是按时间保存的,同时评价也没有考虑新信息因素. 其中 206 249 对用于训练,76 678 对用于测试.

评价标准:用 DCG 评价排序模型的质量^[8],

$$\text{DCG}_n \text{ 定义为: } \text{DCG}_n = \sum_{i=1}^n \frac{G_i}{\log_2(i+1)}.$$

i 是文档列表中的位置, G_i 是相关度函数. 另外一个度量标准是 NDCG, 定义为: $\text{NDCG}_n = Z_n \sum_{i=1}^n \frac{G_i}{\log_2(i+1)}$. Z_n 是标化因子,当 NDCG 理想化列表时,该值为 1.

$\text{DCG}_{\text{regular}}$, 针对常规排序数据,表示全局的相关性,不需要考虑新信息的一致性,也不需要弱化非新信息相关性.

$\text{DCG}_{\text{nodemote}}$, 针对新信息排序数据,表示与非新信息相关性.

$\text{DCG}_{\text{demote}}$, 针对新信息排序数据,表示与新信息和非新信息相关性.

经过编程,执行程序,得到模型的性能指标,见表 1.

从表 1 可以看出,就 $\text{NDCG}_{5, \text{regular}}$, 复合模型的性能优于基础模型,这是因为相关性值作为一个新信息特征,显著地改进了新信息相关性结果. 常规测试数据集下,权重模型给出最好的 NDG 值. 新信息

测试数据集中,迁移学习模型对常规数据集下的 DCG 值有很大的改进.

表 1 模型性能比较表

模型	NDCG ₅ , regular	NDCG ₅ , demote	NDCG ₅ , nodemote
基础模型	0.650 7(0.0%)	0.807 6(0.0%)	0.807 9(0.0%)
复合模型	0.696 4(6.8%)	0.802 7(-0.5%)	0.806 7(-0.2%)
权重模型(1:1)	0.713 4(9.6%)	0.810 3(0.2%)	0.813 5(0.7%)
权重模型(1:2)	0.712 6(9.4%)	0.804 6(-0.2%)	0.810 6(0.2%)
权重模型(1:3)	0.710 2(9.0%)	0.800 6(-0.8%)	0.805 8(0.3%)
迁移学习模型	0.676 2(3.7%)	0.876 3(8.4%)	0.886 4(9.6%)

6 结束语

本文提出了一个搜索敏感新信息的模型,将从 Web 页面中,通过监控一定时间区间内查询条件的数量和文档增长的数据,判断哪些信息属于敏感新信息.利用新信息的多个时间特征,结合传统搜索引擎下根据检索条件得到的检索结果,基于集成的思想,实现包含新信息数据的敏感新信息排序算法,解决新信息数据不足的问题.实验表明,该模型既保证常规信息的相关性又很好地实现了对新信息排序.

参考文献:

[1] Baeza-Yates R, Saint-Jean F, Castillo C. Web dynamics, age and page quality[C]// String Processing and information Retrieval. Mideplan, Chile; IEEE, 2002: 453-461.

[2] Berberich K, Vazirgiannis M, Weikum G. Time-aware authority rankings[J]. Internet Math, 2005, 2(3): 301-332.

[3] Li X, Liu B, Yu P. Time sensitive ranking with application to publication search [C] // Proceedings of Eighth IEEE International Conference on Data Mining.

[s. l.]; IEEE, 2008:893-898.

[4] Pasca M. Towards temporal web search[C]// proceedings of the 2008 ACM symposium on Applied computing. NY, USA; ACM, 2008:1117-1121.

[5] Arikan I, Bedathur S, Berberich K. Time will tell:Leveraging temporal expressions in[C]// proceedings of the second ACM International Conference on web Search and Data Uining Breaking Results; ACM, 2009.

[6] Zhang R, Chang Y, Zheng Z, et al. Search result re-ranking by feedback control adjustment for time-sensitive query[C]// Proceedings of the North American Chapter of the Association for Computational linguistics-Human Language Technologies (NAACL HLT) Proceedings of Stroudsburg, PA, USA; IEEE, 2009: 165-168.

[7] Konig A C, Gamon M, Wu Q. Click-through prediction for news queries[C]// Proceedings of the 32nd Annual ACM SIGIR Conference on Research and Development in formation Retrieval. [s. l.]; ACM, 2009: 347-354.

[8] Jarvelin K, Kekalainen J. Cumulated gain-based evaluation of it techniques[J]. ACM Transactions on information systems, 2002(20):422-446.

作者简介:

王 文 男, (1974—), 讲师, 硕士研究生. 研究方向为信息检索与智能信息处理.

王树锋 男, (1968—), 博士, 副教授. 研究方向为智能信息处理.

朱锡芳 男, (1965—), 博士, 教授. 研究方向为光学工程、应用电子学.

(上接第 112 页)

[8] ADI 公司. AD9751(10-Bit, 300 MSPS High Speed TxDAC+ D/A Converter)[DB/OL]. [2013-01-20]. http://www.analog.com/static/imported-files/data_sheets/AD9751.pdf.

[9] TI 公司. TPS79333(超低噪声、高 PSRR、快速射频、高电平启用的 200mA 低压降稳压器)[DB/OL]. [2013-1-20]. <http://www.ti.com/cn/cn/lit/ds/sym-link/tps79333.pdf>.

[10] TI 公司. CDCM61002(Two Output, Integrated VCO, Low-Jitter Clock Generator)[DB/OL]. [2013-01-20]. <http://www.ti.com/cn/cn/lit/ds/sym-link/cdm61002.pdf>.

[11] Boris Drakhlis. Calculate oscillator jitter by using phase-noise analysis. Microwaves & RF[R]. 2001:82-157.

作者简介:

文德钢 男, (1988—), 硕士研究生. 研究方向为无线通信和信号链技术.

林基明 男, (1970—), 博士, 教授, 博士生导师. 研究方向为无线通信技术和超声电子学.

樊孝明 男, (1971—), 硕士, 讲师. 研究方向为射频微波电路、高速数字信号传输和无线通信技术.