时间感知的 Web 搜索研究

林 盛[1)] 金培权[1),2)] 赵旭剑[3)] 岳丽华[1),2)]

[1)](中国科学技术大学计算机科学与技术学院 合肥 230027)

[2)](中国科学院电磁空间信息重点实验室 合肥 230027)

[3)](西南科技大学计算机科学与技术学院 四川 绵阳 621010)

摘 要 如何利用时间信息改善 Web 搜索效果是近年来的一个研究热点,这是因为大多数的 Web 网页都包含有时间信息,同时许多 Web 查询也含有时间查询词.文中围绕时间感知的 Web 搜索需求,重点研究了两个方面的问题,即查询时间词扩展和时间感知的搜索结果排序,提出了基于查询词和时间词共现关系的查询时间词扩展算法,以及结合了文本相关度和时间相关度的时间感知排序算法.作者建立了一个时间感知的 Web 搜索原型系统,并在大规模真实数据集上进行了实验.实验结果表明作者提出的算法在搜索效果上有明显的改善,并且具有较好的时间性能.

关键词 时间词扩展;Web 搜索;时间感知排序

中图法分类号 TP311 **DOI 号** 10.11897/SP.J.1016.2015.02163

Research on Time-Aware Web Search

LIN Sheng[1)] JIN Pei-Quan[1),2)] ZHAO Xu-Jian[3)] YUE Li-Hua[1),2)]

[1)](*School of Computer Science and Technology, University of Science and Technology of China, Hefei* 230027)

[2)](*Key Laboratory of Electromagnetic Space Information, Chinese Academy of Sciences, Hefei* 230027)

[3)](*School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang, Sichuan* 621010)

Abstract Utilizing time information to improve the effective of Web search has been a research focus in recent years, because most Web pages contain time information and many Web queries involve time-related query terms. Motivated by the needs of time-aware Web search, in this paper, we concentrate on two key issues, namely temporal keyword expansion and time-aware ranking. Particularly, we propose a new algorithm for temporal word expansion which is based on the co-occurrence of query terms and time words. In addition, we present a new time-aware ranking algorithm considering both textual relevance and time relevance. We build a prototype system and conduct experiments on real data sets to evaluate the performance of our proposal. The results show that the proposed algorithms are time efficient and have a substantial improvement on the effective of Web search.

Keywords temporal word expansion; Web search; time-aware ranking

1 引 言

时间在许多研究领域都扮演着很重要的角色,例如信息抽取、话题检测、问答系统、查询日志分析以及 Web 搜索等.在网页中时间信息主要以时间表达式的形式出现,它们可以分成两种类型,即显式时间(如“2013 年 5 月 20 日”)和隐式时间(如“今天”).时

收稿日期:2013-11-09;最终修改稿收到日期:2015-01-06.本课题得到国家自然科学基金面上项目(60776801,61379037)、中国科学技术大学研究生科技创新与社会实践资助专项和中国科学技术大学出国研修基金项目资助.**林 盛**,男,1987 年生,博士研究生,主要研究方向为 Web 时空信息抽取及检索.E-mail:linsh@mail.ustc.edu.cn.**金培权**(通信作者),男,1975 年生,博士,副教授,中国计算机学会(CCF)高级会员,主要研究方向为移动对象数据库、时空数据库、面向新型硬件的数据库技术、Web 信息抽取与检索.E-mail:jpq@ustc.edu.cn.**赵旭剑**,男,1984 年生,博士,讲师,主要研究方向为话题动态演化、信息检索.**岳丽华**,女,1952 年生,教授,博士生导师,主要研究领域为数据库系统及其应用、信息集成、实时数据库.

间表达式的不同展现形式以及 Web 网页存在的大量时间信息使 Web 搜索领域的研究产生了几个难题：

(1)查询时间词扩展问题：如何确定一个查询的时间查询词？这一问题的主要挑战在于很多时候用户在进行 Web 搜索时并不会主动提供显式的时间信息.因此如何挖掘用户的时间查询需求，并将时间查询需求以时间词的方式扩展到用户查询中，这是时间感知 Web 搜索中需要重点研究的一个问题.

(2)时间感知的搜索结果排序问题：如何将网页中的时间信息整合到 Web 搜索结果的排序中？搜索结果的排序是决定 Web 搜索效果的关键因素.在时间感知的 Web 搜索中，不仅需要考虑查询中的时间信息[1-2]，还要考虑网页中的时间信息，并通过考虑查询与网页之间的时间相关性来提升时间相关的 Web 搜索效果.

针对第 1 个问题，一种方法是对大量用户的查询日志进行分析，得到和查询相关的时间词进行扩展，但这种方法对于最新的事件查询效果较差，因为新的时间在查询日志中的查询相对较少.针对第 2 个问题，即时间感知的排序，难点在于如何确定时间之间的相似度.由于查询和文档中都可能包含多个时间表达式，时间相似度就是需要考虑每一对查询和文档的时间表达式之间的相似度，其中还需要考虑时间信息之间的相交关系和时间词本身的重要程度.已有的研究大多数只考虑了网页的创建时间，往往难以准确地反映网页的时间信息.

本文围绕时间感知的 Web 搜索需求，主要研究上述的查询时间词扩展问题和时间感知的搜索结果排序问题.为了解决并验证以上问题，我们实现了一个时间感知搜索的原型系统(Time-Aware Search Engine，TASE)[3].该系统的基本架构如图 1 所示.

该系统集成了时间感知查询和查询时间词扩展的模块，可以为时间感知搜索的相关工作提供实验平台.本文所提出的在这篇文章中，基于上述搜索原型系统我们对多个算法进行了实验.本文的主要贡献可以总结为以下几点：

(1)提出了一个对查询进行查询时间词扩展的方法，该方法利用查询词和时间词在文档中的共现关系来确定和查询相关的时间词，这样对于未提供时间词的查询可以给用户推荐一些相关时间限制词.

(2)针对时间文本查询我们提出了一个改进的时间感知排序算法，该算法结合了查询和文档之间的文本相关度和时间相关度，能够更好地评价查询与网页的相关性.

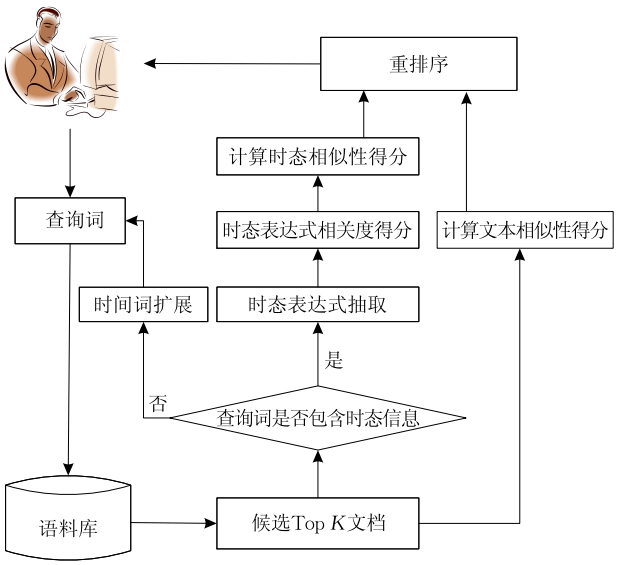

图 1 TASE 的实现架构

(3)实现了一个时间感知搜索的原型系统，并在该系统上基于真实数据集开展了对比实验，验证了所提出的算法的有效性.

2 相关工作

时间感知排序是利用查询和文档中潜在的时间意图来改进 Web 搜索的结果，时间感知排序方面的研究主要包括时间信息的抽取和利用该时间信息来回答有时间依赖的查询.

最近的研究表明在 Web 搜索中大量的查询中包含时间信息，有些时间关键词是显式的时间表达式[1]，有些时间信息是查询中潜在的时间语义[2].Li 和 Croft[4]把时间信息整合到语言模型中(Time-Based Language Model)，其实现方式就是给文档赋予一个先验的概率，该先验概率是一个随着文档发布时间的变长而指数递减的函数，他们旨在解决面向最近查询的问题，以至于最近发布的文档具有更高的相关概率.Diaz 和 Jones[5]也利用文档的创建时间来分析搜索引擎返回的候选文档集合的时间概率分布，根据这样的概率分布来确定这个查询潜在的时间语义，他们的实验表明结合查询的潜在的时间信息和文档的内容能够提高检索的平均准确率.Kanhabua 和 Nørvåg[6]利用语言模型来标注查询的潜在时间词，在他们的方法中，3 种不同的方法来确定一个查询可能的时间，然后通过考虑不确定性和不考虑不确定性的模型对时间查询结果进行重排序，把和查询包含的时间语义最相关的文档的相关度得分进行提高.这些方法的一个主要局限性就是

他们只考虑了文档的发布时间，而不考虑文档的内容时间，这样就忽略了文档中包含的大量的有价值的时间信息.

文献[1,7]中考虑了网页的内容时间. Berberich 等人[1]把时间信息整合到查询似然的语言模型中，它考虑了查询和文档的时间信息的不确定性，也就是说查询和文档的时间信息在有交集的时候两者就有时间相似性权重，而不需要两者完全相等时才有时间相似性权重，在这一方面我们的工作和他们的工作具有相同点，即我们都把网页的内容时间引入到排序中来改善时间相关检索的性能. Li 等人[7]提出的方法是把网页中每个关键词都赋予一个和它最相关的时间词，然后基于这样的一个关键词-时间词的配对集合，他们提出了一个时间增强的语言模型. 在这些研究中，所有网页的内容时间相对于网页来说都被认为有一致的权重，即他们没有去区分确定网页中哪一个时间对于网页来说更加的重要.

与前面介绍的研究不同的是，Metzler 等人[2]提出了一个直接从查询日志中挖掘和查询关键字词相关的时间信息的方法，他们的方法并没有去分析网页的发布时间和内容时间. 这里他们限定和查询关键词相关的时间信息是以年为粒度的，而对于其他的时间粒度并没有进行支持，利用查询日志的方法的另外一个缺陷就是对于新网页来说，查询日志里包含的信息比较好，不能做很好的预测.

在文献[8-10]中，Campos 等人根据搜索引擎返回的网页快照进行分析，通过抽取并分析网页快照片段中的时间信息，对查询结果进行分类，另外，他们还分析了这些时间信息和查询词之间的相关性，用于识别用户的隐式时间意图，他们只是利用了搜索引擎返回的网页片段，而没有对网页全文进行分析，从而丢失了许多有用的信息，而且他们只考虑了年这个时间粒度，对于其他粒度的时间信息未予考虑.

3　查询时间扩展

在本章中，我们将介绍在用户进行查询时查询时间词扩展的相关技术和实现，要进行查询时间词的扩展是因为在很多时候 Web 用户在进行查询操作的时候，用户可能并不知道他所需要查询的确切的时间区间或时间点，也就是说用户不能够显式的提供和查询相关的查询时间词，这里我们将会介绍如何在文档中挖掘潜在的查询相关的查询时间词.

3.1　候选句子的获取

在文档中，我们认为要是有一个关键词和某些时间词相关的话，那么在海量文档中该关键词一般都会和这些时间词在同一个句子中一起出现. 因此，在时间词扩展的时候我们主要针对包含时间词的句子进行处理，得到和查询词相关的时间词. 针对一个查询 Q，我们将按照图 2 的流程得到候选句子的集合 S. 首先，我们在原始的文档集中对查询 Q 进行查询，得到 Top-k 个初始的查询结果，然后对这 k 个文档进行时间词抽取以及分句的处理，保留包含时间词的句子，得到候选的句子集合 S，如式(1)所示.

$$S=\{S_1,S_2,S_3,\cdots,S_n\} \tag{1}$$

$$S_i=\langle\{t\text{-}temp\},\{t\text{-}text\}\rangle \tag{2}$$

集合 S 中每个句子由时间词和文本关键词组成，如式(2)所示，其中，$t\text{-}temp$ 表示句子中的时间词集合，$t\text{-}text$ 表示句子中普通关键词的集合.

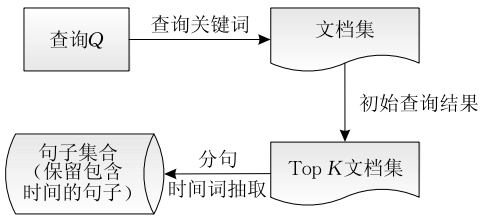

图 2　候选句子获取流程

3.2　权重矩阵

在本节中，我们将定义两个权重矩阵，即词项-句子矩阵 $\boldsymbol{A}$ 和时间-句子矩阵 $\boldsymbol{B}$，如图 3 所示. 词项-句子矩阵 $\boldsymbol{A}$ 用来表示每一个词和每一个句子之间的权重关系，它和词项在句子中出现的频率和句子的新鲜度相关，句子的新鲜度是指句子中的时间词离当前时间越近，句子越新鲜. 时间-句子矩阵 $\boldsymbol{B}$ 用

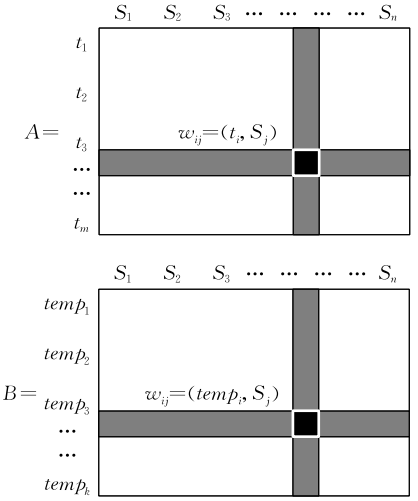

图 3　权重矩阵

来表示每一个时间词和每一个句子之间的权重关系，它和时间词的类型（显式时间、隐式时间）以及时间词的粒度（年、月、季度等）相关.

如式(3)～(5)所示，对于词项-句子矩阵 $\boldsymbol{A}$，w_{ij} 表示词项 t_i 在句子 S_j 中出现的频率和句子 S_j 的权重的乘积. 而句子权重 w_{S_j} 根据句子的长度，句子所在文档的发布时间来决定.

$$w_{ij}=tf_{t_i}\times w_{S_j} \tag{3}$$

$$w_{S_j}=f_{len}(S_j)\times f_{dct}(S_j)=f_{len}(S_j)\times e^{\lambda\times|t_c-t_{s_j}|} \tag{4}$$

$$f_{len}(S_j)=\begin{cases}1, & \text{若 } length(S_j)>20\\ 0.5, & \text{其他}\end{cases} \tag{5}$$

其中，tf_{t_i} 表示词 t_i 在句子 S_j 中出现的频率，t_c 表示当前的时间，离当前时间越近的 $f_{dct}(S_j)$ 的值越大，则句子的权重越大，而且我们认为长句子（句子的长度大于 20）具有较高的权重.

如式(6)～(9)所示，对于时间-句子矩阵 $\boldsymbol{B}$，w_{ij} 由句子 S_j 中的时间词 $temp_i$ 的时间类型和时间粒度来决定. 即

$$w_{ij}=f_{\text{occ}}(temp_i,S_j)\times f_{\text{type}}(temp_i)\times f_{\text{gran}}(temp_i) \tag{6}$$

$$f_{\text{occ}}(temp_i,S_j)=\begin{cases}1, & \text{如果 } temp_i \text{ 在 } S_j \text{ 中出现}\\ 0, & \text{否则}\end{cases} \tag{7}$$

$$f_{\text{type}}(temp_i)=\begin{cases}1, & \text{如果 } temp_i \text{ 是显式时间}\\ 0.5, & \text{否则}\end{cases} \tag{8}$$

$$f_{\text{gran}}(temp_i)=\begin{cases}1, & \text{如果 } temp_i\in\{\text{天，月}\}\\ 0.75, & \text{如果 } temp_i\in\{\text{季度，年}\}\\ 0.5, & \text{如果 } temp_i\in\{\text{年代}\}\end{cases} \tag{9}$$

以上几式表明如果时间词 $temp_i$ 不在句子 S_j 中出现，则其权重 w_{ij} 为 0，并且显示时间词的权重比隐式时间词的权重高，时间粒度为天和月的时间词和其他时间粒度的时间词的权重要高.

3.3 扩展时间词的确定

这里，如式(10)所示我们可以将一个查询 Q 表示成一个由词项集合组成的向量形式.

$$\boldsymbol{Q}=[w_{t_1},w_{t_2},w_{t_3},\cdots,w_{t_m}] \tag{10}$$

如果查询 Q 包含词项 t_i，则 $w_{t_i}=1$，如果查询 Q 不包含词项 t_i，则 $w_{t_i}=0$. 根据以上的相关定义，我们可以根据式(11)计算与查询 Q 相关的时间词的权重.

$$\begin{aligned}\boldsymbol{Q}_{temp}&=[w_{temp_1},w_{temp_2},w_{temp_3},\cdots,w_{temp_k}]\\&=\boldsymbol{Q}\times\boldsymbol{A}\times\boldsymbol{B}^{\mathrm{T}}\end{aligned} \tag{11}$$

这里 w_{temp_i} 表示时间词 $temp_i$ 和查询 Q 的相关性，w_{temp_i} 值越大，表明时间词 $temp_i$ 和查询 Q 的相关性越大. 因此，可以根据 w_{temp_i} 的值对时间词进行降序排列，排在前面的几个时间词可以作为查询 Q 的扩展查询时间词，可以让用户选择选用哪个时间词进行查询，也可以结合这些扩展的时间词进行再次检索，把时间相关的查询结果返回给用户.

4 时间感知的排序

我们首先介绍文档的时间表示模型，然后讨论时间相似度算法，最后，基于时间相似度计算查询与文档的时间相关性并结合文本相关性对网页搜索结果进行重排序.

4.1 时间表示模型

一个时间表达式或者一个文档的发布时间可以根据式(12)表示成一个四元组[1].

$$T=(tb_l,tb_u,te_l,te_u) \tag{12}$$

其中 tb_l 和 tb_u 分别表示一个时间区间的开始时间点的下界和上界，相应地，te_l 和 te_u 分别表示该时间区间的结束时间点的下界和上界. 例如，时间表达式“2013 年 3 月”可以表示为(2013/03/01,2013/03/31,2013/03/01,2013/03/31).

一个文档 d 由文本部分 d_{text} 和时间部分 d_{time} 组成，其中 d_{time} 由文档的发布时间 $PubTime(d)$ 和文档的内容时间 $ContentTime(d)$ 组成，$ContentTime(d)$ 可以表示成 $\{t_1,\cdots,t_k\}$.

一个时间查询 q 由文本关键词 q_{text} 和时间关键词 q_{time} 组成. 在时间感知的检索过程中存在两种模式：(1) 包含模式(inclusive)和 (2) 不包含模式(exclusive). 对于包含模式，查询的文本关键词 q_{text} 由查询中的文本词和时间词共同组成，对于不包含模式，查询的文本关键词 q_{text} 只由文本词组成，而不包含时间词. 因此，对于一个用户输入查询“航天飞船 1981 年 4 月 12 日”，在包含模式中 q_{text} 表示为{航天飞船 1981 年 4 月 12 日}，而在不包含模式的情况下表示为{航天飞船}.

4.2 时间相似性

我们的方法的表示基于论文[1]中提出的方法，在我们的方法中，我们结合了网页中时间信息和网页的相关度的得分. 在这里我们先介绍一下他们的方法，在他们的方法中考虑了两种模型，一个是不考虑时间不确定性的模型(Uncertainty-Ignorant Language Model，LMT)，LMT 模型中查询时间和文档时间对的相关度计算如式(13)所示.

$$P(t_q|t_d)_{\text{LMT}}=\begin{cases}0, & t_q\neq t_d\\ 1, & t_q=t_d\end{cases} \tag{13}$$

其中 $t_d\in ContentTime(d)$，只有当 t_q 和 t_d 完全相等

的时候相似度得分才为 1，即满足条件$(tb_l=qb_l)\wedge(tb_u=qb_u)\wedge(te_l=qe_l)\wedge(te_u=qe_u)$时才为 1.

另一个模型是考虑不确定性的模型(Uncertainty-Aware Language Model，LMTU)，LMTU 模型中考虑了查询时间词 t_q 和文档时间词 t_d 的所有可能情况的交集. 因此 LMTU 中查询时间和文档时间的相关度计算如式(14)和(15)所示，其中在文献[1]中详细地阐述了$|T|$的快速的计算方法.

$$P(t_q|t_d)_{\mathrm{LMTU}}=\frac{|t_q\bigcap t_d|}{|t_q|\cdot|t_d|} \tag{14}$$

其中 $t_d\in ContentTime(d)$，$|t_q\bigcap t_d|$可以表示为

$$|t_q\bigcap t_d|=(\max(tb_l,qb_l),\min(tb_u,qb_u),\max(te_l,qe_l),\min(te_u,qe_u)) \tag{15}$$

则时间相关度 $S''(q_{\mathrm{time}},d_{\mathrm{time}})$可以根据式(16)通过计算查询中每一个时间表达式和文档中每一个时间表达式之间的相关度得分的综合得到.

$$S''(q_{\mathrm{time}},d_{\mathrm{time}})=\prod_{t_q\in q_{\mathrm{time}}}P(t_q|d_{\mathrm{time}})=\prod_{t_q\in q_{\mathrm{time}}}\left(\frac{1}{|d_{\mathrm{time}}|}\sum_{t_d\in d_{\mathrm{time}}}P(t_q|t_d)\right) \tag{16}$$

基于上面的时间相似性计算公式，我们引入我们之前工作[11-13]中的方法来确定网页中的时间表达式和网页的相关度得分 $Score(T)$，这样在计算时间相似性的时候就能区分不同时间对于网页的重要性，在我们的方法中，我们在计算查询时间表达式 t_q 和文档中时间表达式 t_d 相关度的时候采用式(17)和(18)的方法进行计算.

$$P(t_q|t_d)_{\mathrm{ELMT}}=\begin{cases}0, & 若\ t_q\neq t_d\\ Score(t_d), & 若\ t_q=t_d\end{cases} \tag{17}$$

$$P(t_q|t_d)_{\mathrm{ELMTU}}=Score(t_d)\cdot\frac{|t_q\bigcap t_d|}{|t_q|\cdot|t_d|} \tag{18}$$

在上面的公式中，t_q 是查询中 q_{time} 中的一个时间，t_d 是文档中 d_{time} 中的一个时间，$Score(t_d)$ 是时间 t_d 和文档 d 的相关度得分，分数越高表示时间 t_d 对于文档 d 越重要. ELMT(Enhanced LMT)是一个不考虑时间不确定性的模型，相应的 ELMTU(Enhanced LMTU)是一个考虑时间不确定性的模型.

4.3　网页的重排序

为了使搜索引擎具有时间感知的能力，需要整合查询和网页的时间相关度的得分来确定一个查询和一个网页的最终的相关度得分，根据该得分对网页进行重排序. 之前的研究[1,14]的一种方案就是把时间相似性整合到相应的语言模型中，文献[15-16]把通过混合模型线性的将时间相似性引入到概率模型和向量空间模型中. 在本文中，我们把原始的文本相似性得分 $S'(q_{\mathrm{word}},d_{\mathrm{word}})$ 和时间相似性得分 $S''(q_{\mathrm{time}},d_{\mathrm{time}})$ 进行线性求和，在线性求和之前文本相似性得分和时间相似性得分都要进行归一化，例如，所有的相似性得分都除以 Top K 个文档中相似性得分最高的分值. 对于给定的时间查询 q，文档 d 将根据式(19)的得分公式进行重排序.

$$S(q,d)=(1-\alpha)\times S'(q_{\mathrm{word}},d_{\mathrm{word}})+\alpha\times S''(q_{\mathrm{time}},d_{\mathrm{time}}) \tag{19}$$

其中参数 α 来调节文本相似性和时间相似性的重要程度.

5　实验与分析

5.1　实验设置

我们的实验数据集来源于新浪网新闻网页和纽约时报网站上抓取的真实的数据. 新浪网数据集包含了 2012 年全年和 2013 年第一季度的新闻网页集合，总共包含 198360 个网页. 纽约时报包含 1812933 个网页，网页的发布时间在 1981 年到 2011 年之间. 其中包含各个领域的新闻文章，如商业，体育等. 我们利用这个数据集来构建简单的检索系统，在本文中，我们使用 Apache Lucene 3.5.0 来构建索引，其中原始的文本相似性可以由 Lucene 提供的函数得到.

在这篇文章中，我们实现了 6 种不同的时间感知的排序算法，即 LMT、LMTU、ELMT、ELMTU、TS[6] 和 TSU[6]，其中前面 4 种算法在第 4 节中已经进行了介绍，这里我们将对 TS 和 TSU 算法进行简要的介绍.

TS 算法没有考虑时间的不确定性，因此 $P(t_q|t_d)_{\mathrm{TS}}$ 采用和 $P(t_q|t_d)_{\mathrm{LMT}}$ 一样的计算方法，但是 t_d 对应的是网页的发布时间，而不是网页的内容时间. 和 TS 算法一样，TSU 利用的也是网页的发布时间，另外它还考虑了时间的不确定性，$P(t_q|t_d)_{\mathrm{TSU}}$ 使用如式(20)和(21)所示的一个指数衰退函数来计算.

$$P(t_q|t_d)_{\mathrm{TSU}}=DecayRate^{\lambda\cdot\frac{|t_q-t_d|}{\mu}} \tag{20}$$

$$|t_q-t_d|=\frac{|tb_l^q-tb_l^d|+|tb_u^q-tb_u^d|+|te_l^q-te_l^d|+|te_u^q-te_u^d|}{4} \tag{21}$$

其中 $t_d=PubTime(d)$，$DecayRate$ 和 λ 是常数，$0<DecayRate<1$ 并且 $\lambda>0$，μ 是一个时间单元的长度，其中心思想就是 t_q 和 t_d 的相似性得分随着它们

之间的距离增大而减小,距离越小,时间相似性越高.

5.2 查询时间扩展实验结果与分析

我们利用新浪网的数据集来验证我们查询时间扩展的算法的有效性.首先使用 Lucene 对该数据建立索引,对于查询 Q,我们利用文本相关性得到前 50 个网页集合,根据本文的算法处理得到这些网页中包含时间词的句子集合.根据文献[4]中的实验我们在算法中句子新鲜度的参数 λ 取值为 -0.02.我们选取了常用的事件词作为查询关键词,如表 1 所示,我们列出了部分的查询关键词.

表 1 查询时间扩展算法查询词示例

查询关键词	查询关键词
地震	日全食
洪水	踩踏
强拆	禽流感
总统大选	伦敦奥运会

在表 2 中我们用"地震"为例展示了我们查询时间扩展算法计算得到的前 10 个查询时间扩展词,这 10 个查询时间扩展词中有 9 个是和地震这个事件相关的时间词,其中一个不正确是由于网页中会穿插一下新闻报道的时间以及同一个句子中有时会出现两个相对独立的事件造成的.

表 2 查询时间扩展算法示例

时间扩展词	时间扩展词权重
2013 年 2 月 19 日	6.371291376493909
2012 年 8 月 12 日 18 时 47 分	5.216149412392835
2012 年 8 月 31 日 20 时 47 分	4.260718944831057
2012 年 9 月 1 日	4.260718944831057
2013 年 2 月 18 日	3.920794693227021
2013 年 2 月 19 日当天	3.920794693227021
2012 年 9 月 22 日	2.556431366898634
2013 年 1 月 19 日 22 时 56 分	2.0
2013 年 1 月 9 日	2.0
2013 年 1 月 9 日 15 时 26 分 04 秒	2.0

我们对所有选取的查询关键词的查询时间扩展词的前 10 个结果进行统计扩展词的准确率,表 3 中统计了这些查询时间扩展词在 $P@1$、$P@5$ 和 $P@10$ 上的平均准确率,实验结果表明我们的算法能够取得较高的准确率,对于用户选取所需的查询时间词具有较大的帮助.要是对于更大的数据集,和事件比较相关的时间词就能出现得比较多,小部分不太符合规范的句法对结果的影响也会随之减小,实验的结果会更好一些.

表 3 查询时间扩展算法准确率

$P@k$	准确率
$P@1$	1.00
$P@5$	0.90
$P@10$	0.87

5.3 时间感知排序实验结果与分析

在这一节中,我们将讨论不同时间感知排序算法的性能,所有的排序算法都使用同样的查询和同样的数据集,即我们抓取的纽约时报的数据集.我们选取了 30 个时间敏感的查询,表 4 列出了其中部分的查询.

表 4 时间感知算法查询词示例

文本关键词	时间关键词
kurt cobain	April 5, 1994
mickey mouse	1930s
dallas mavericks	June 2011
obama	November 4, 2008
boston red sox	October 27, 2004
michael jackson	1982
muammar qaddafi	October 20, 2011
iphone 4	June 7, 2010
thomas edison	1891

在我们的原型系统中,我们使用包含模式(inclusive)并且只考虑文本相似性的算法作为基准算法,表示为 TFIDF,其中文本相似性在我们的方法中使用由 Lucene 得到的文本相似性得分.针对每一个查询,我们使用 7 种不同的算法(TFIDF,ELMT,ELMTU,LMT,LMTU,TS 和 TSU)以及后 6 种算法对应的两种查询模式("包含模式"和"不包含模式")对搜索结果进行重新排序,我们对每种情况下在排序结果前 20 个文档进行相关性判定.我们采用 $P@5$、$P@10$、$P@20$ 和 MRR 来评价排序性能.

在我们的实验中存在多个可调节的参数,包括用于重排序的候选文档集的大小、TSU 算法的参数以及文本相似性得分和时间相似性得分的权衡因子 α.我们比较了用于重排序的不同文档候选集大小(50、500 和 1000)对实验结果的影响,实验结果表明用于重排序的文档候选集为 1000 时能取得较好的实验结果,因此在我们的系统中针对每种算法我们都对前 1000 个搜索结果计算他们的文本相关性和时间相关性,然后进行重排序.TSU 算法的参数采用和原算法的参数一致,即 $DecayRate=0.5$,$\lambda=0.5$ 和 $\mu=6$ 个月.而权衡因子 α,我们针对 10 种不同的取值(0.1,0.2,…,1.0)进行比较,当 α 取值为 0.1 时表示我们几乎不关注时间相似性,主要由原始的文本相关度来决定重排的结果,相反地,当 α 的取值为 1.0 的时候表明我们在重排的时候仅仅考虑时间相似性.

图 4 和图 5 显示了用 $P@5$ 评价方法不同的算法随着 α 取值的变化时排序性能的变化.$P@10$、

$P@20$ 和 MRR 的结果和 $P@5$ 类似. 可以看出，不同的算法基本上都有一个共同点，即当 α 在一定的取值范围内检索的性能会随着 α 变大而变大，当 α 的值超过了一定的值后，检索性能会随着 α 的增大而变小，这就说明了对于时间感知的查询过多或过少的考虑时间相似性都会给检索性能带来负面的影响.

表 5 显示了不同算法在最优参数下的检索结果. 从表中可以看出，所有的时间感知排序方法都比不考虑时间相似性的基准算法的结果要好，我们的方法(ELMT，ELMTU)对比于其他的算法在大多数的情况下比其他的算法性能好，这表明区分网页中时间表达式的重要程度有利于提高时间相关检索的检索性能.

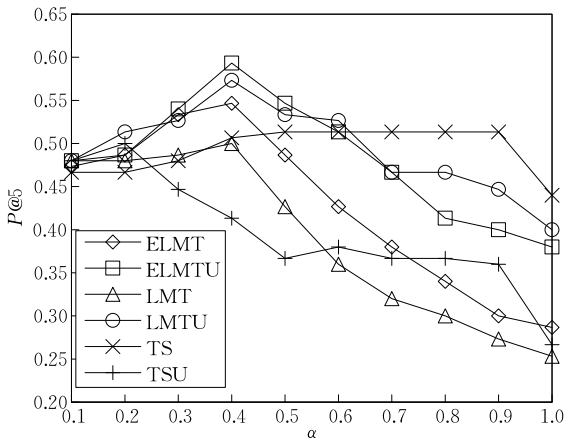

图 4　在包含模式下 $P@5$ 随权衡参数 α 的变化

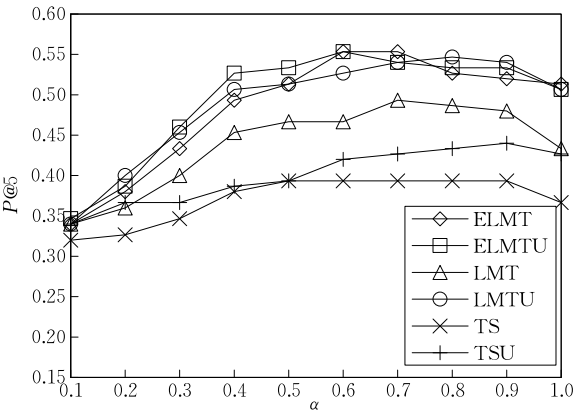

图 5　在不包含模式下 $P@5$ 随权衡参数 α 的变化

表 5　不同排序算法的检索性能

算法	检索模式	$P@5$	$P@10$	$P@20$	MRR
TFIDF	inclusive	0.467	0.417	0.383	0.682
ELMT	inclusive	0.547	0.460	0.418	0.754
ELMTU	inclusive	**0.593**	**0.513**	**0.450**	**0.801**
LMT	inclusive	0.500	0.427	0.390	0.751
LMTU	inclusive	0.573	0.500	0.447	0.796
TS	inclusive	0.513	0.447	0.397	0.738
TSU	inclusive	0.500	0.463	0.380	0.701
ELMT	exclusive	**0.553**	**0.513**	0.467	**0.757**
ELMTU	exclusive	**0.553**	0.503	0.472	0.746
LMT	exclusive	0.493	0.413	0.373	0.718
LMTU	exclusive	0.547	0.503	**0.482**	0.735
TS	exclusive	0.393	0.347	0.300	0.603
TSU	exclusive	0.440	0.390	0.357	0.632

我们可以看出，实验中的各个时间感知的算法的准确率还不够高，一方面是由于本实验中使用从纽约时报网站中抽取的新闻网页作为数据集，网页的数量相对于互联网的海量资源还有很大的差距，搜索返回的结果中相关网页的个数是比较有限的；另外，我们实验中是对初始的根据文本相关度返回的搜索结果进行重排序的，在这一步中可能会漏掉一些相关的文档，这对时间感知排序算法的准确率也有一定的影响.

6　结束语

时间是信息的重要维度，已成为 Web 搜索近几年研究的一个重点方向. 本文研究了时间感知 Web 搜索中的两个关键问题：查询时间词扩展和时间感知的搜索结果排序. 我们提出了一种基于查询词和时间词在文档中共现关系的查询时间词扩展算法，以及一种结合文本和时间相关性的搜索结果排序方法. 为了验证算法性能，我们构建了一个时间感知搜索的原型系统，在该系统上我们实现了多个时间感知的查询排序算法，并基于真实的新浪网新闻网页上进行了实验. 实验结果表明，我们的查询时间词扩展算法具有较高的准确率，能够给用户提供准确可用的和查询关键词相关的时间词. 另外，在纽约时报的数据集上进行的实验表明了我们提出的考虑网页中时间词对于网页具有不同重要程度的想法能够改进时间感知排序算法的有效性.

在下一步的工作中，我们将着重考虑如何使时间信息更好地结合到搜索引擎的其他模块中，例如网页的索引. 此外，我们还将研究时间信息在其他 Web 相关领域中的应用，例如 Web 网页的自动摘要、问答系统、个性化推荐系统等.

参　考　文　献

[1] Berberich K, Bedathur S, Alonso O, Weikum G. A language modeling approach for temporal information needs// Proceedings of the 32nd European Conference on Information Retrieval (ECIR). Milton Keynes, UK, 2010: 13-25

[2] Metzler D, Jones R, Peng F, Zhang R. Improving search relevance for implicitly temporal queries//Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). Boston, USA, 2009: 700-701

[3] Lin S, Jin P, Zhao X, Yue L. TASE: A time-aware search engine//Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM). Maui, USA, 2012: 2713-2715

[4] Li X, Croft W B. Time-based language models//Proceedings of the 12th ACM International Conference on Information and Knowledge Management (CIKM). New Orleans, USA, 2003: 469-475

[5] Diaz F, Jones R. Using temporal profiles of queries for precision prediction//Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). Sheffield, UK, 2004: 18-24

[6] Kanhabua N, Nørvåg K. Determining time of queries for re-ranking search results//Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries (ECDL). Glasgow, UK, 2010: 261-272

[7] Li X, Jin P, Zhao X, et al. NTLM: A time-enhanced language model based ranking approach for Web search//Proceedings of the WISE 2010 Workshops-WISE 2010 International Symposium WISS, and International Workshops CISE, MBC. LNCS 6724. Hong Kong, China, 2010: 156-170

[8] Campos R, Dias G, Jorge A, Nunes C. GTE: A distributional second-order co-occurrence approach to improve the identification of top relevant dates in Web snippets//Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM). Maui, USA, 2012: 2035-2039

[9] Campos R, Jorge A, Dias G, Nunes C. Disambiguating implicit temporal queries by clustering top relevant dates in Web snippets//Proceedings of the 2012 IEEE/WIC/ACM International Conferences on Web Intelligence (WI-IAT). Macau, China, 2012: 1-8

[10] Campos R, Dias G, Jorge A. What is the temporal value of Web snippets?//Proceedings of the WWW 2011 Workshop on Linked Data on the Web (TWAW). Hyderabad, India, 2011: 9-16

[11] Lin S, Jin P, Zhao X, et al. Extracting focused time for Web pages//Proceedings of the 13th International Conference on Web-Age Information Management (WAIM). Harbin, China, 2012: 266-271

[12] Zhao X, Jin P, Yue L. Automatic temporal expression normalization with reference time dynamic-choosing//Proceedings of the 23rd International Conference on Computational Linguistics (COLING). Beijing, China, 2010: 1498-1506

[13] Lin S, Jin P, Zhao X, Yue L. Exploiting temporal information in Web search. Expert Systems with Applications, 2014, 41(2): 331-341

[14] Dakka W, Gravano L, Ipeirotis P. Answering general time-sensitive queries. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(2): 220-235

[15] Jin P, Li X, Chen H, Yue L. CT-rank: A time-aware ranking algorithm for Web search. Journal of Convergence Information Technology, 2010, 5(6): 99-111

[16] Kanhabua N, Nørvåg K. A comparison of time-aware ranking methods//Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). Beijing, China, 2011: 1257-1258

LIN Sheng, born in 1987, Ph. D. candidate. His research interests include spatiotemporal information extraction and Web search.

JIN Pei-Quan, born in 1975, Ph.D., associate professor. His research interests include moving objects databases, spatiotemporal databases, databases on new hardware, Web information extraction and retrieval.

ZHAO Xu-Jian, born in 1984, Ph.D., lecturer. His research interests include dynamic topic evolution, information retrieval.

YUE Li-Hua, born in 1952, professor, Ph. D. supervisor. Her research interests include database system and application, information integration, real-time database.

Background

Time-aware Web search is a promising way to improve the effectiveness of Web search by making use of the time intention in queries and the temporal contents in documents. Recent researches on Web search have shown that a majority of Web queries contain explicit or implicit time words. To improve the effectiveness of searches on time-based Web queries, researchers have proposed some methods. Some of them proposed to incorporate time into a language model by

assigning a priority of time to each document. Such a value of priority can be computed in terms of an exponential decay function on documents' creation dates. Based on the priority of time, time-related queries can be better evaluated and fresh documents are more possible to get high rankings. Previous studies have been also focused on the time information embedded in the contents of Web pages, i. e. , content time. They extracted the content time expressions in Web pages and then normalized them into some unified form. However, they did not consider the different impacts of different content time expressions in reflecting a document's time information.

In the research area of temporal keyword expansion, previous studies used documents' creation dates to measure the distribution of retrieved documents as well as to create the temporal profile of a query. Some researchers employed language models to associate time with documents. One limitation in these studies was that they only considered create dates and ignored content time in the contents of documents. Another method for temporal keyword expansion was mining temporal patterns directly from query logs. However, this method was not suitable for new queries because they were usually not listed in query logs.

In this paper, we proposed a time-aware search prototype system and implemented several baseline algorithms in the system. We proposed a new time-aware ranking algorithm which considered the text similarity and temporal similarity and made use of the temporal expressions' importance of Web pages. We also proposed a new temporal keywords expansion algorithm, which was based on the co-occurrence between query words and temporal expressions in Web pages.

This work is supported by the National Science Foundation of China under Grant Nos. 60776801 and 61379037, the USTC Special Grant for Postgraduate Research, Innovation and Practice, and the OATF project Funded by University of Science and Technology of China.