基于中文新闻网页主题划分的未来预测系统

王晨光，王崇文，邴杰
（北京理工大学软件学院，北京 100081）

摘要：当今，互联网上有很多与未来相关的信息，但信息量大，且易变，使得用户查询未来信息时显得力不从心，因为用户无法辨别这些信息的正确性，所以无法选出最佳、最可靠的答案。使用基于中文网页，尤其是中文新闻网页的未来信息检索方法，检测时间实体；采用主题划分的方法，分析用户查询，获取网页集；通过多模式匹配算法过滤网页集，并将最终结果呈现给用户。实验结果表明，系统采用上述方法后，能够正确、稳定、高效的返回给用户可靠、满意的未来信息。

关键词：人工智能；未来信息检索；事件预测；时间信息分析；查询子主题分析

中图分类号：TP393

Future-related Future Prediction System by Query Subtopic Analysis Based on Chinese News Web Pages

Wang Chenguang, Wang Chongwen, Bing Jie
(School of Software,Beijing Institute of Technology,Beijing 100081)

Abstract: Nowadays, there is lots of information related with the future, which are large in quantity but inflexible, users feel hard to find the future-related information correctly. To find the temporal entity, using future-related information retrieval method based on Chinese web pages, especially for the Chinese news web pages; analyzing user' query and achieving web page collections using the subtopic analysis method; using the multi-pattern matching method to process the web page collections, also present the final results to users. The experimental results show the method can correctly, stably, efficiently return the reliable, satisfied future-related information to users.

Keywords:artificialintelligence;future-relatedinformationretrieval;eventprediction;temporalinformation analysis;query subtopic analysis

0 引言

互联网时代，已经有很多与未来相关的信息[1]，这些信息容易得到。人们时常想组织以及安排他们的生活，于是写出了一些文字性的材料，包括计划、期待发生的事情以及有关未来的预测等等，这些材料都与未来特定的事件对象有关。M. D. Choudhury[2]，Y. Liu[3]做过关于特殊领域未来信息的研究，但是没有涉及应用于普遍领域的研究工作。

针对以上问题我们提出——基于中文新闻网页主题划分的未来预测。问题涉及从中文新闻网页中抽取时间信息，并且将这种时间信息与标准的全文检索相结合，从而回答查询。所以，查询既包括时间信息，又包括普通文本信息。查询通常包括一个或更多主题[4]。划分主题的方法对于预测系统的效率有很大的帮助，避免了一些传统搜索引擎因不区分主题而导致的与查询不相关结果出现的问题。我们定义的检索方法将查询分成若干子主题（假设的查询只有一个主题），所得到的检索结果要比通过传统方法检索得到的结果更加准确、可靠。该方法首先从查询中抽取子主题；然后，对于每一个子主题生成一个子查询，扩展子查询后，

基金项目：国家科技部“863”计划项目：基于 SOA 数字农业智能决策服务关键技术研究（2007AA10Z234）
作者简介：王晨光（1989-），男，硕士生，自然语言处理. E-mail: 2007270702@bit.edu.cn

逐个子查询获取检索结果；接着综合评价子主题的重要性来对结果集排序；最后通过系统后处理生成最终结果。

1 中文网页主题划分的未来信息检索分析

1.1 处理概要

首先，从查询中抽取出子主题，并且为子主题加上时间属性；然后针对每一个子主题生成相应的子查询；接着扩展子查询；然后，检索所有包括相应子主题、以及相应时间属性的网页；最后，经过后处理，生成最终网页结果集。

数据来源于百度的中文新闻网页，并且使用特定的网页片段作为系统处理的对象以及最终结果呈现的单元。

当我们分析存在于新闻网页中的未来信息时，需要注意到很多与时间相关的问题[5]：第一，网页创建时间需要考虑。因为这个时间决定网页是否有可能包括未来信息。如图 1 所示，网页 1 的创建时间是在 2008 年，网页中所提及的时间是在 2009 年，我们把后者称作网页的关注时间，如果我们的阅读时间是在 2010 年，那么在阅读此网页时，网页就没有包含任何未来信息；与之不同的是，网页 2 的关注时间是 2012 年，所以包括未来信息。

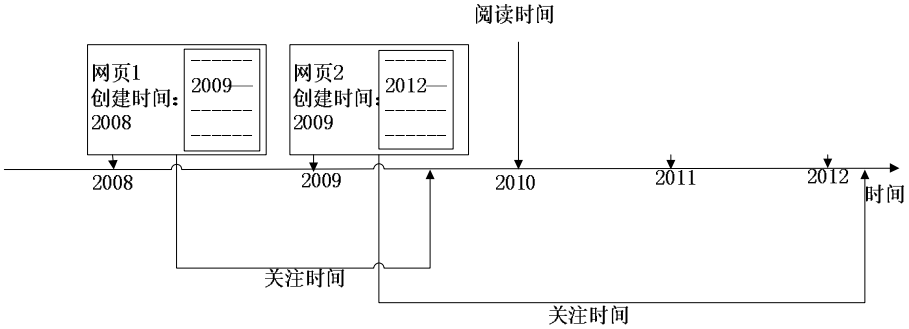

图 1 使用三种时间的示意图

Fig. 1 Visualization of three notions of time

一般情况下，我们将与网页相关的时间分成以下三种：网页创建时间；网页关注时间；网页阅读时间。

网页中的时间短语可以大致划分为两类：第一类是绝对时间，另外一类就是相对时间。前者指的是那些可以直接从网页中被提取出来的时间点、时间段。例如，7 月 22 日，2012 年等等。而后者指的是这样的一些时间短语（两年后、明年），可以利用一些绝对时间进行相对时间的精确定位。系统利用网页的创建时间来将相对时间换算成为绝对时间。

时间短语拥有不同的粒度值。例如，以天为单位，以月为单位，以季节为单位，还有以年为单位等等。简单起见，选择以年为单位作为时间粒度。

置信度是必须考虑的问题，任何与未来事件相关的信息都有置信度。我们在置信度方面所做的约定如下：未来事件发生的概率：数据集中事件出现的次数与数据集中所有未来信息数量的比值。未来事件发生的时间：网页中所有时间的中间时间。系统比较的是相同的事件，而不是具有相同发生时间的事件。网页创建时间是系统最终结果生成的前提，并且创建较晚相对创建较早的网页中描述的未来事件更加真实、正确，应当给予更高的置信度。

1.2 处理流程

步骤 1- 子主题抽取： 任何一个查询都包括一个或者更多的主题[6]。长一些的查询可能

包括更多子主题。子主题定义如下：从整个查询中抽取出的一段文本。抽取的方法以及文本的选取方法有很多，具体的选择视目的以及查询的组成而定。

步骤 2-子主题查询短语的抽取：为每个子主题创建子查询，系统在检索时使用子查询。子查询是从每一个子主题中抽取出的短语集。

步骤 3-子主题的重要性评价：每一个从查询中抽取出的子主题都被赋予相应的重要性值。步骤 7 将重要性值综合利用，决定结果集的排序。1.4 节中，将详细描述子主题重要性的计算方法。

步骤 4-基于未来信息的查询扩展：使用百度新闻的 API 来搜集针对查询的结果集。针对每一个子主题对应的子查询，子查询短语集被发送给百度新闻的 API，得到的网页片段包括新闻的标题，以及整个原始网页内容中包括子查询的部分。其中，网页片段是针对子查询的检索结果的摘要，网页片段在后文中都简称为网页。从百度新闻 API 返回的数据包含大量的、重复的事件，是由于在一段时间内各大新闻门户网站长期报道，或是特定时期的大众特定关注所致。百度新闻 API 返回的结果集中总会有与子查询密切相关的重要信息以及有用信息被忽略、丢失。

为了得到更加合理的结果集，我们将子查询转化为一系列被时间短语约束的查询。系统在检索前，在子查询中加入时间信息，这样就使得百度新闻 API 返回的结果集中的事件都在所限制的时间范围之内。查询的约束建立在具备唯一性的时间框架之下，时间框架的起始时间为 2010 年 6 月，结束时间为 2013 年 6 月，以 6 个月为时间单位，所以生成了 6 个时间段约束的子查询，完成了子查询扩展。时间段约束的子查询被发送给百度新闻 API，就能收集到 30 条检索后返回的网页（每个时间段返回前 5 个网页）。图 2 展示了如何将原始的子查询分成一系列时间段约束的子查询：

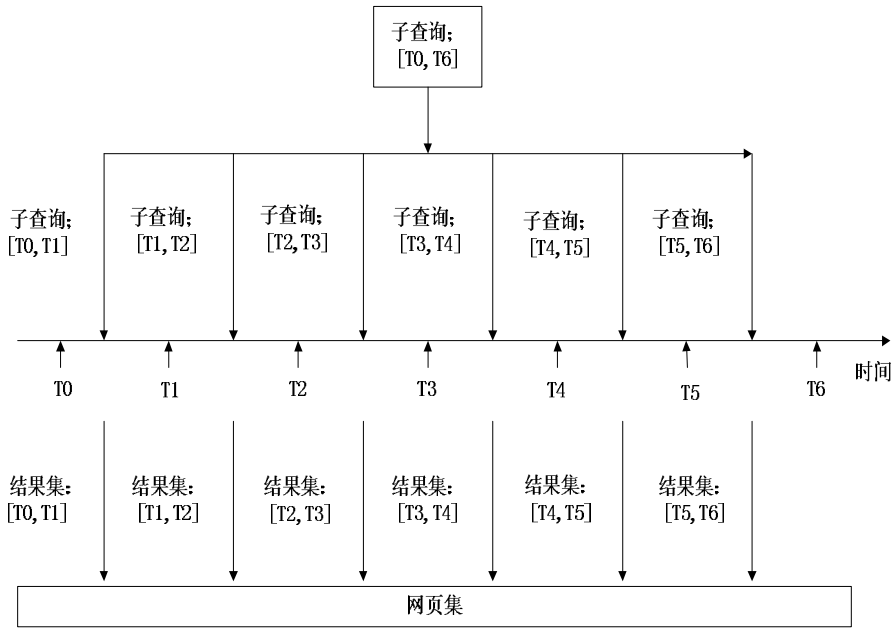

图 2 时间短语约束子查询示意图

Fig.2 Subtopic query reformulation into a series of temporarily constrained subtopic queries

使用上述方法能够得到一个数量足够的网页集作为结果集。然而，还是无法提供充足的未来信息资源，为了得到更加丰富的未来信息资源，也就是得到更多的新闻网页，进一步进

行子查询扩展。向子查询中增添表征未来的时间短语，例如，“2014 年”。扩展后的子查询包括原始的子查询短语以及添加的时间短语。针对每一个子查询，添加“2011 年”到“2020 年”共 10 个时间短语，相当于 1 个子查询变为了 10 个查询，然后对于每一个扩展后的子查询，使用上述的时间段约束方法，将每一个子查询置于时间框架下，每一个被扩展后的子查询返回 30 个结果，就能够得到 300 条结果，数量上符合系统的设计要求，图 3 表示完整的子查询扩展过程：

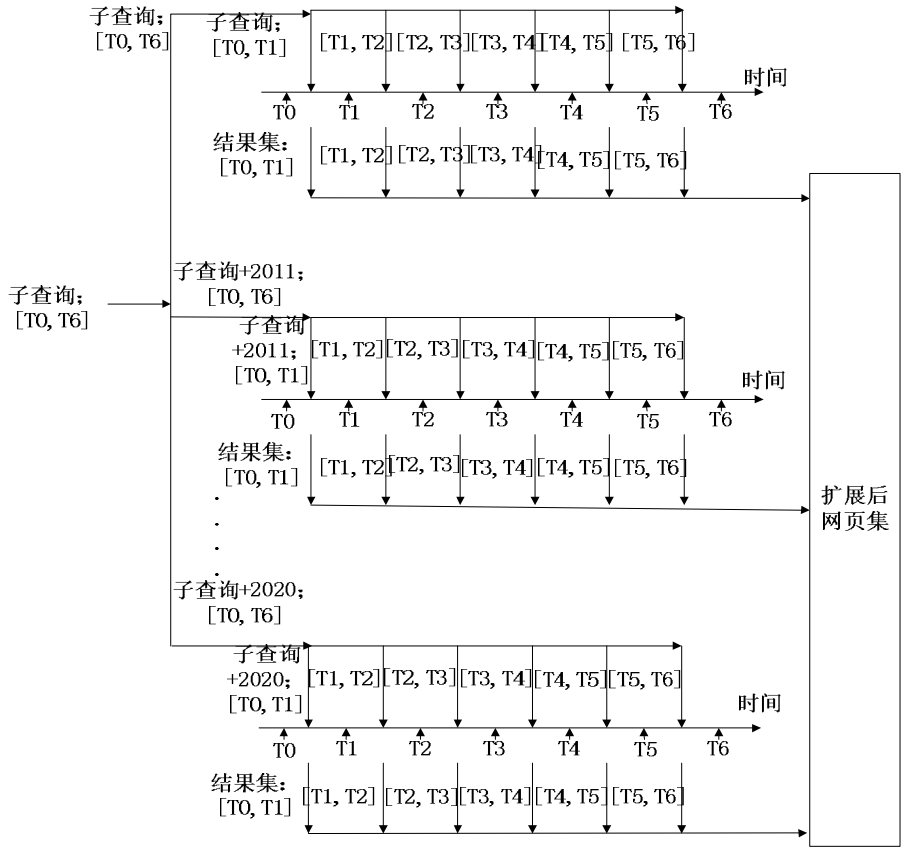

图 3 子查询扩展示意图

Fig.3 Visualization of subtopic query expanding

步骤 5-网页关注时间的检测：首先，检索新闻网页中的时间短语，目的是检索得到表示未来的时间短语。然后，将相对时间转化成为绝对时间，所采用的方法是将检测得到的网页创建时间当作固定的时间点，去转化相对时间。例如，网页的创建时间如果是“2010 年”，那么网页中出现的“15 年后”就会被当成“2025 年”进行后续的处理，也就是当作此网页关注时间。某个网页检测过后只得到一个时间短语，此即为网页关注时间；某个网页中的时间短语很多时，此网页的关注时间取中间值，例如，有 3 个时间短语，“2010 年 9 月”，“2010 年 7 月”，“2012 年 8 月”，认为“2011 年 7 月”为此网页关注时间。

步骤 6-基于子主题的排序：使用一种传统的排序模型，处理针对于每一个子查询而得到的网页集，为每一个子主题生成一个已经排好序的网页集。模型在 1.3.2 节介绍。

步骤 7-网页集集成：子主题的重要性作为准则，集成所有使用子主题扩展后所搜集得到的网页集，生成一个针对原始查询的未来事件列表。模型在 1.3.1 节介绍。

步骤 8-最终结果集生成：最后一步将对上一步生成的列表进行进一步的过滤，以避免与查询无关的结果被最终呈现。所使用的算法将在 1.5 节介绍。

系统整体处理流程见图 4：

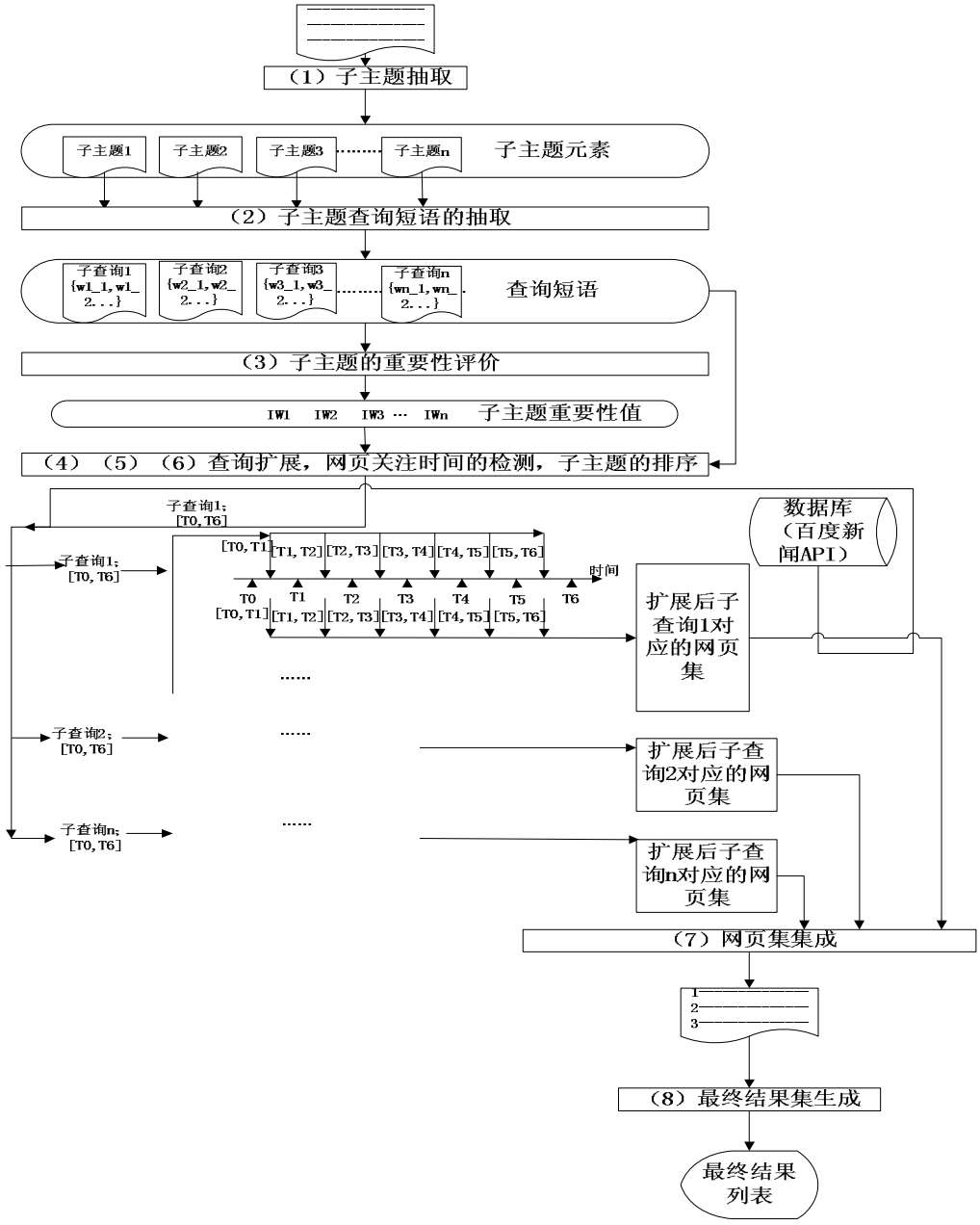

图 4 系统处理流程示意图
Fig.4 Visualization of system processing

1.3 检索模型

1.3.1 检索结果集成模型

中文新闻网页 D 与查询 Q 的相关性为 $Score(D,Q)$。集成检索模型：

$$Score(D,Q)=\sum_{i=1}^{m}(Subscore(D,SQ_i)*IW_i) \qquad (1)$$

m 是子主题的数量；针对每一个子主题的子查询 SQ_i，它是由从第 i 个子主题中抽取出

的一系列查询短语构成的集合；$Subscore(D,SQ_i)$ 是中文新闻网页 D 与子查询 SQ_i 之间的相关性；IW_i 是第 i 个子主题的重要性值。

1.3.2 中文新闻网页检索模型

使用最佳匹配检索模型来计算 $Subscore(D,SQ_i)$，实验中，使用 Okapi BM25 模型[7]：

$$Subscore(D,SQ_i)=\sum_{T\in SQ_i}\omega\frac{(k_1+1)tf}{K+tf}\frac{(k_3+1)qtf}{k_3+qtf} \quad (2)$$

SQ_i 是一个包括短语 T 的查询；ω 是 SQ_i 中短语 T 的 Robertson/Sparch Jones 权重值；k_1、k_3 是参数，均为常量；K 可以表示为：$K=k_1((1-b)+b\frac{dl}{avdl})$；$tf$ 表示短语 T 在中文新闻网页 D 中出现的频率；qtf 是短语 T 在 SQ_i 中出现的频率；dl 以及 $avdl$ 分别表示 D 的期望长度，以及在所有网页集中网页的平均期望长度。

1.4 子主题重要性计算

子主题重要性是由子主题中查询短语特殊性来综合衡量的。当查询短语特殊性高时，说明查询短语仅仅出现在有限数量的子主题中；相反，当查询短语的特殊性低时，说明查询短语出现在很多的子主题中。在计算子主题特殊性时使用熵，熵在系统中的应用如下：

一个随机变量的熵 n_j，表示查询短语 ω_j 出现在每一个子主题所对应的查询短语集：$\{SQ_1,SQ_2,.....,SQ_{m-1},SQ_m\}$ 中的频率：

$$n_j=-\sum_{i=1}^{m}P_j^i\log_2 P_j^i \quad (3)$$

m 是查询中所包括的子主题数量；P_j^i 是查询短语 ω_j 出现在子主题 SQ_i 中的概率，估算 P_j^i 的值使用 SQ_i 中 ω_j 的出现频率 $tf_{j,i}$，还用到了 ω_j 在所有子主题中出现的总频率。所以 P_j^i 的计算公式如下(4)：

$$P_j^i=\frac{tf_{j,i}}{\sum_{k=1}^{m}tf_{j,k}} \quad (4)$$

将上述(3)，(4)合并：

$$n_j=-\sum_{i=1}^{m}\frac{tf_{j,i}}{\sum_{k=1}^{m}tf_{j,k}}\log_2\frac{tf_{j,i}}{\sum_{k=1}^{m}tf_{j,k}} \quad (5)$$

当查询短语 ω_j 仅仅出现在一个子主题当中时，n_j 为 0。一种更加理想的频率计算方法能使得结果更加光滑，我们将 δ 加入上式的每一个 $tf_{j,i}$ 以及 $tf_{j,k}$ 后面，(5)变形为：

$$n_j=-\sum_{i=1}^{m}\frac{tf_{j,i}+\delta}{\sum_{k=1}^{m}(tf_{j,k}+\delta)}\log_2\frac{tf_{j,i}+\delta}{\sum_{k=1}^{m}(tf_{j,k}+\delta)} \quad (6)$$

以上，查询短语 ω_j 在所有子主题中出现的总频率使用熵来衡量，ω_j 的重要性 S_j：

$$S_j=\log_2\sum_{i=1}^{m}tf_{j,i}-n_j \quad (7)$$

子主题中所包括的查询短语重要性的和看作是每一个子主题的重要性，所以，使用子主题中短语的数量对子主题的重要性计算进行归一化处理：

$$IW_i = (\frac{1}{\log(1+|SQ_i|)}\sum_{\omega_j \in SQ_i} s_j) * \exp[\frac{\mathrm{Re}adTime - Date(Doc)}{\mu + 1}] \qquad (8)$$

$Date(Doc)$ 是网页创建时间； μ 是用以决定网页的创建时间对于子主题重要性的影响的参数，默认为 2。

1.5 中文网页内容多模式匹配过滤算法

系统使用主题划分方法检索后得到的网页集要进行后处理，才能生成将最终结果。我们借鉴了多模式匹配算法[8]。

将算法改进，以更好的引入系统，简介如下：

1）问题描述：多模式算法问题描述如下：已知：中文字母表 C，夹杂字母表 Σ，以及针对查询的子查询集 $\{SQ_1, SQ_2, \ldots\ldots, SQ_{m-1}, SQ_m\}$；求解：查找不符合查询所有子主题的结果，及其在网页集中的位置。

2）算法实现：输入参数：网页集，用于匹配的模式数组；输出结果：存放子查询短语出现位置的 2 维数组 *D[][] ／C[][]*。其中，第 1 维用于指明是查询短语数组中的第几个，第 2 维用于存放该查询短语在结果集中第几次出现。

3）使用方法：最后将出现在 2 维数组 *D[][] ／C[][]*第 1 维的结果删除。

2 实验结果

评估结果的方法：首先人工标注一些实验结果；然后人工分析，确定一组值，当作结果的基线；最后，用这个底线去评价系统给出的其他结果。

使用 20 个查询作为实验样本确定系统的基线，如表 1 所示；然后人工分析返回的结果，以及结果中的所有显示参数；最终，我们认为满足以下三个条件的结果是满意的：能从结果中分析出查询的主题；结果中给出的未来事件与查询有关，并且是非常重要的事件；事件的实际发生时间与系统给出的发生时间一致。

表 1 实验样本
Tab.1 Test sample

实验样本				
伦敦奥运会	奥巴马	周杰伦	非诚勿扰	石油危机
经济危机	大学生就业	NBA	世界杯	北京
房地产	低碳	波兰	南非	美国
春运	阿凡达	世界末日	迈克杰克逊	惠普

系统给出的查询“奥林匹克运动会”的结果见表 2：

分析下面实验结果，排在前两位的都是有关 2016 年夏季奥运会的问题，巴西的里约热内卢最终获得了主办权，而另一大竞争对手西班牙的马德里的新闻排在第 4 的位置也就十分合理了；伦敦奥运会颁布会徽是势在必行的事情，所以可能性以及热度评分均较高，排在第 2 也是十分正确的；虽然 2014 年的国际青年奥运会在南京举办，但是由于仅仅是国际青年奥运会的第二届比赛，所以热度不高，综合排序到第 3 位也是合理的； 夏季特奥会的影响力不大，但是也是奥林匹克大家庭的组成部分，排在第 5 位依旧合理。

表 2 查询“奥林匹克运动会”的结果
Tab.2 Result of query “Olympic”

标题	关键词	发生时间	可能性	热度评分
里约热内卢市将举办 2016 年奥运会	中国、竞技、北京、奥运会	2016 年	93.7%	9.3
2012 年伦敦奥运会会徽	会徽、巴萨罗那、伦敦、评估	2012 年	94.1%	9.2
2014 年国际青年奥林匹克运动会	南京、青年、火炬、组织者	2014 年	91.4%	8.8
国际奥委会新闻	马德里、奥委会、国际	2016 年	92.1%	9.0
2011 年夏季特奥会	运动员、希腊、世界、华盛顿	2011 年	90.2%	8.0

综合分析后，系统给出的结果是合理的，并且满足了上述三个满意度条件。

3 结论

我们经常有这种需要，想知道一些即将发生的，以及一些发生可能性较大的未来事件，来帮助自己规划未来，或是了解一些焦点事件的发展趋势等等。例如，我们想知道“2012 年是否是世界末日”，我们很关注、可预见、有理可依的未来事件信息。本文介绍了针对上述需求所做的研究工作以及研究成果。我们提出了一整套解决方案用以解决上述需求所涉及的难题。现阶段，“基于中文新闻网页主题划分的未来预测系统”开发完成，系统使用未来信息检索方法检测时间实体；使用主题划分方法分析查询，结合时间实体信息，来获取网页集；最后使用了多模式匹配的方法过滤网页集，并且生成最终结果集列表。

4 致谢

这项研究工作的顺利开展以及完成要特别感谢北京理工大学薛静峰教授给予的大力支持以及悉心指导。

[参考文献] (References)

[1] R. Baeza-Yates. Searching the Future. Proceedings of ACM SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval (MF/IR 2005), 2005.
[2] M. A. Hearst: Multi-Paragraph Segmentation of Expository Text, In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, pp.9–16, 1994.
[3] M. D. Choudhury, H. Sundaram, A. John and D. D. Seligmann. Can Blog Communication Dynamics be Correlated with Stock Market Activity? Proceedings of the 19th ACM Conference on Hypertext and Hypermedia, pp.55-60, 2008.
[4] Y. Liu, X. Huang, A. An and X. Yu. ARSA: a Sentiment-aware Model for Predicting Sales Performance Using Blogs. Proceedings of the 30th Annual International ACM SIGIR Conference, pp. 607-614, 2007.
[5] Adam Jatowt, Kensuke Kanazawa, Satoshi Oyama and Katsumi Tanaka. Supporting Analysis of Future-related Information in News Archives and the Web, Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2009), ACM Press, Austin, USA, pp. 115-124 (2009).
[6] Toru Takaki, Atsushi Fujii, Tetsuya Ishikawa. Associative document retrieval by query subtopic analysis and its application to invalidity patent search. CIKM 2004: 399-405.
[7] S.E. Robertson and S. Walker. Okapi/keenbow at TREC-8, In Proceedings of the Eighth Text Retrieval

Conference (TREC-8), NIST Special Publication 500-246, pp.151 – 161, 2000.

[8] Zhou Xueguang , Zhang Huanguo. Flexible Pattern Matching Algorithm in Chinese Strings. Proceedings of the 27th Chinese Control Conference July 16-18,2008,Kunming,Yunnan, China.