基于浅层分析的网页相关度研究*

昝红英　苏玉梅　孙斌　俞士汶
北京大学计算语言学研究所 北京 100871
E-mail: zanhy@pku.edu.cn

摘　要　本文介绍了北京大学天网知名度系统的设计与开发工作，重点论述了其中网页相关度评价的因素、算法和相应的检索结果。系统在北京大学天网搜索引擎的基础上，运用中文信息提取的新技术，结合网页信息的特点，针对名人网页的检索提出了一种新的网页相关度评价算法，改善了检索结果的排序合理性，提高了名人网页检索服务的质量。

关键词　相关度，检索服务，信息提取，特征信息

The WebPages Relevance Research Based on the Shallow Parsing

ZAN Hongying　SU Yumei　SUN Bin　YU Shiwen
The Institute of Computational Linguistics, Peking University, Beijing, China, 100871
E-mail: zanhy@pku.edu.cn

Abstract　This paper introduced the design and implementation of Tianwang Fame System. It mainly discussed on the factors and algorithms that affect matching of a named entity with webpages' relevance evaluation. Based on the Tianwang Search Engine of Peking University, the Fame System adopted new techniques in Chinese information extraction according to the features of webpage information. The paper proposed a new method to the relevance evaluation of webpages against attributes of named enties. This method optimizes the order of the search results, and improves the service quality of Tianwang Fame System.

Keywords　relevance, searching service, information extraction, feature information

*本文的研究工作得到国家自然科学基金项目(69973005)、863 项目（2001AA114040） 和北大-IBM 创新研究院项目的支持。

一、引言

在信息技术迅猛发展的今天，因特网确实是一个信息的宝库，但很大程度上它还只是信息的堆砌，因此它更像一个宝矿，等待人们去挖掘。然而，目前的网络服务例如搜索引擎等还远没有达到人们的要求，常常是没有语义分析，只是根据用户所给出查询词串的逻辑组合机械地找出一系列匹配网页，造成垃圾信息过多。实际上，对于网上海量信息的检索，人们更看重的是准确性、及时性，希望能得到个性化的检索服务，希望将他们最关心的网页排在前面。但是用户关心什么内容？怎样将搜来的网页按照用户检索的相关程度排序？一直是网络服务提供者特别是搜索引擎开发者研究的课题。

北京大学计算机系网络实验室 1997 年推出了 “天网”，在国内中文搜索引擎方面享有很高的知名度，特别是天网 FTP 检索在国内首屈一指。2001 年网络实验室又推出“燕穹”中国 Web 信息博物馆，为国人了解中文网站的历史提供了可能。据“天网” 搜集的网页估计，中文（简体）网页数已超过 1 亿，网上海量信息的涌现迫使国人越来越依赖于搜索引擎，因此，提高网上信息检索的智能性，按照用户关心焦点的相关度排序检索结果，提供个性化检索服务已势在必行。本项目在天网搜索引擎的基础上，利用中文信息提取的先进技术，对中文（简体）网页的资源进行再次挖掘，以大众关心焦点之一的名人网页作为起点，尝试网页的个性化检索服务，力求提高网络服务的质量。

二、相关研究现状

1. 搜索引擎技术

目前，搜索引擎对用户查询的返回结果一般都是按某种顺序进行排序的，只是不同的搜索引擎的排序方式不同，实现技术不同。

Google 是目前公认为最好的搜索引擎，它采用的技术是基于网络的超级链接结构，通过被链接的次数来判断网页级别而排序检索结果，即 pagerank 方法。该方法是不针对特定主题的，它根据整个 web 之间的链接关系来确定每个网页的重要程度，从总体上而言，Google 的搜索结果排序反映的是网页被链接的次数，即一个网页的重要程度取决于其它网页（或网站）对它的评价，因此不能满足每个查询者个人对查询主题的具体要求，对网上广告等链接信息也不能轻易予以剔除。

天网（Tianwang）搜索引擎将搜来的原始网页进行一系列处理：首先是中文分词，对其中的实词进行词频、位置及标记的分析，形成关键字、位置和权值等信息组成的网页表示库，进而建立索引库，保证搜索出的结果与用户的查询串相完全匹配或部分匹配，搜索结果实现了一定程度的合理排序。由于对网上新词的识别率较低，加之权值计算等其它一些原因，天网的网页搜索不如其文件搜索效果好，仍需进一步改进。

2. 信息提取技术

对网页采用基于内容的分析或/和基于链接分析的方法进行相关度评价，使检索结果排序合理，是每个搜索引擎追求的目标。可以说，所有的搜索引擎，都采取一定的技术来争取检索的准确性、完整性和及时性。但是，目前的技术现状仍不能满足人们的检索需求。

本文我们利用部分中文信息提取技术来对网页内容进行浅层分析，希望增加对网页相关度评价的准确性。按照 MUC-7(Message Understanding Conferences)的任务规定，一个完整的信息提取过程包括如下 5 个（由简到难的）阶段：

- 命名实体 NE (Named Entities): 提取文本中相关的命名实体，包括人名、机构/公司名称的识别
- 实体关系 ER (Entity Relations): 提取命名实体之间的各种关系（事实）等，例如 Location_of, Employee_of, Product_of 等关系
- 模板脚本 Template Scenario (Event Structures): 提取指定的事件，包括参与这些事件中的各个实体、属性或关系
- 共指 Coreference (Identity descriptions): 代词、名词共指分析
- 模板合并 Template Merger: 把相同的事件合并成为一个

根据名人网页检索的特点，在天网知名度系统中我们主要利用了信息提取中较为简单成熟的命名实体和实体关系的识别技术。

三、天网知名度系统中对名人网页的相关度评价

1. 名人网页的相关度界定

天网知名度服务质量的核心是对名人网页的相关度评价，它直接决定了对名人网页检索结果的排序。这里的相关度，是一个比较主观、比较宽泛的概念，指某个网页与用户查询关键词（主题描述）的相关程度。到底怎样才是相关的呢？我们认为，搜索引擎的服务对象是用户，查询结果的呈现应该以人为本。名人网页的相关度以用户的需求为第一判断标准。对名人网页，即使对同一名人的相关网页，不同的用户也会有不同的关心焦点。

2. 影响名人网页相关度的因素

在天网知名度系统中，我们将用户关心的名人称为实体。针对名人网页的特点，要求用户将其要查询的名人信息分类予以注册，系统将为每个用户登记专用的实体信息，形成个人信息库和实体信息库，以保证尽量满足每个用户的个性化检索需求。

每个实体的信息分为以下八类：

- 名人所在的领域（政府、科教、业界、影视等）
- 名人的名字，包括别名、笔名、艺名等，保证检索的完整
- 名人所在的工作单位，以下各信息作为识别名人的重要标志
- 名人的职业描述（主席、书记、教授、记者、演员等）
- 名人的兼职单位（可以有多个）

- 名人的社会形象
- 特征词（用户关心的特征描述）
- 名人的代表作（著作、作品名、产品名等）

以上信息基本涵盖了名人网页的基本特征，系统将根据用户注册的这些信息去分析过滤每个网页，计算网页的相关度。

另外，网页与一般纯文本的区别是它含有丰富的标记信息，用于不同格式的显示，客观上不同程度地吸引人们的注意力，因此不同标记的内容应该赋以不同的权值，对相关度应有不同的贡献。目前绝大多数网页还是 HTML 格式，我们在设计中主要考虑了 HTML4.0 的标记信息，按照重要程度将其分为三类，根据关键词的标记类别累计不同的权值。

3. 名人网页的相关度评价算法

名人网页的相关度评价算法，决定了对名人网页检索结果的排序。在名人网页相关度评价之前，名人知名度系统前期的处理工作有：对天网搜来的原始网页进行标记过滤，中文分词，同时进行人名识别、人名与单位（Employee_of）以及人名与职务（Post_of）等二元关系的识别，形成 BerkeleyDB 形式的网页表示库。网页对名人的相关度用一个 32 位整数表示，所有网页对注册名人的相关度评价结果存在网页相关度评分库中。有了以上的准备工作，名人网页的相关度评价流程如下：

对网页表示库中的每一个网页：

- 检查其人名列表，检索用户信息库，对其中已注册的人名（实体名）建立一个该网页对该人名的相关度评分初值；
- 对检索出的注册名人列表，检查该网页中的二元关系和实体信息库，对符合匹配的关系为该网页的相关度评分增加一定分值，同时利用排除词表过滤掉重名的无关网页；
- 对网页分词中的有效词（对语义理解有效的大部分实词）分别检索实体信息库的八类信息，分不同情况为该网页对名人的相关度评分增加不同分值；
- 对网页分词中的有效词检查其 HTML 标记，分不同情况为该网页对名人的相关度评分增加不同分值；
- 根据网页长度（按词计算）、网页中的人名个数等因素调整其相关度评分值；
- 形成网页相关度评分库。

网页相关度评价模块采用标准 C++编码实现，在 IBM 服务器 Netfinity 7600 PIII（Xeon 700/2M Cache 512MB 4*18.2GB SCSI Hard disk/Ultra2 ）的 Red Hat Linux 7.2 系统下运行正常，对 75 万网页全部处理一遍需要约 80 分钟；另外，系统还实现了对个别网页单独的相关度评价功能，从而保证了系统的及时性。

4. 名人网页检索的实验结果

不同的相关度评价策略对天网知名度检索的结果将有一定影响。我们可以利用的信息有网页的词频、HTML 标记、用户注册的信息以及从网页中提取的二元关系，实验是将这些信息逐步加入我们的评价策略，观察它们检索结果的影响。表中给出了对若干个实体的六种（A--F）评价结果，具体为

A	纯文本(词频)	D	A+二元关系
B	A+结构化用户信息	E	A+B+C
C	A+HTML 标记信息	F	A+B+C+D

我们对每个实体搜集了 20 至 30 个相关网页，并且人工给出了每个网页的相关性评价，分为高[high]、中[mid]、低[low]三个等级，其中遵循的标准（可以根据用户的要求再行调整）为：相关度高的可以是有关目标实体名人的个人信息（生平介绍、作品介绍等），名人领域内的专题报道等；相关度中的可以是名人发表的评论或文章，名人的花边新闻，在其他人的报道中目标实体名人所占份额较多者，或其他不容易归到相关度高或低的情况；相关度低的可以是在其他报道（如新闻）中偶尔提到目标实体人名，引文中偶尔出现目标实体名字，或很多其他人名（编委、会议名单等情况）中偶尔出现目标实体人名等情况。

这里对天网知名度的检索结果与人工评判结果进行计较。基本思想是：在高、中、低三中不同的相关度的网页中，相关度得分应该满足下列关系：

[high] > = [mid]　　[high] >= [low]　　[mid] > = [low]

如果定义已知网页集合中相关度为高、中、低的个数分别

|high|=m;　|mid|=n;　|low|=k

那么总的关系个数为：

Total = m*n + n*k + k*m

评价结果以 x/y 形式给出，其中 y 为有关实体网页中总的关系个数（各个实体总关系不同的原因是我们搜集的网页个数不同），x 为天网知名度系统正确评价的关系个数。

表 不同的相关度评价策略对天网知名度的检索结果

Entity_Name	Result_A	Result_B	Result_C	Result_D	Result_E	Result_F
朱镕基	32/38	29/38	15/38	32/38	32/38	19/38
王刚（中共中央办公厅）	41/44	41/44	29/44	41/44	41/44	38/44
李晓明	55/76	56/76	58/76	56/76	56/76	59/76
迟惠生	66/74	64/74	47/74	61/74	63/74	59/74
王玮	16/18	16/18	15/18	16/18	16/18	15/18
叶天正	57/69	56/69	50/69	55/69	55/69	55/69
杨澜	49/90	48/90	49/90	50/90	52/90	52/90
崔永元	42/98	43/98	39/98	48/98	44/98	43/98
周大新	31/40	32/40	31/40	32/40	33/40	33/40
刘璐	15/15	15/15	15/15	15/15	15/15	15/15
田震	45/66	43/66	45/66	50/66	48/66	48/66
崔健	33/45	30/45	33/45	39/45	35/45	35/45
王刚（主持人、演员）	41/48	41/48	34/48	40/48	40/48	41/48
巩俐	62/96	62/96	62/96	66/96	66/96	66/96
王刚（辽宁足球队）	51/99	51/99	51/99	56/99	54/99	54/99
郝海东	49/69	48/69	49/69	47/69	49/69	49/69

从表中的结果数据可以看出，加入结构化用户信息和二元关系大部分结果有所改进，这说明利用命名实体识别和实体关系提取的浅层分析技术，在网页内容分析中可以进一步尝试。但是，表中也有些不好的结果，特别是将所有的信息逐步加入相关度评价策略是评价的结果并没有呈现明显的单调上升趋势，而是有一定的波动，这需要再实验，调整各种信息的参数。

四、进一步的工作

目前该系统的原型开发已基本完成，并于 2002 年 11 月 27 日顺利通过 PKU-IBM 联合创新研究院管理委员会的中期验收，演示结果得到了 PKU 和 IBM 有关人员的认可。由于只是原型设计，系统还有很大的改进空间，算法还显粗糙，各模块还需进一步完善和优化。

分词与二元关系提取模块是系统处理网页的基础，目前本模块对少数网页处理还有报错现象，对人名识别、网上新词的识别也有待改进；另外，对网页中灵活的描述形式，二元关系的提取仍有不周之处，近期内还需要对关系模版做进一步的调整和扩充工作。

为了增加检索的智能性，我们在本期工作中对实体信息库中出现的词语进行了人工的同义词集概念扩展。从长远计议，我们计划利用由北京大学计算语言学研究所开发与的 WordNet 同构的中文概念语义词典（Chinese Concept Dictionary, CCD）实现对实体信息库中词语的自动概念分级扩展。

另外目前的算法虽然取得了一定的效果，但是算法思想还较为简单粗糙，特征信息还需要人工注册或提取。今后希望实现特征信息在某种程度上的机器学习或自动提取，构建统计与规则相结合的评价模型，并且可以根据领域、用户的选择来调整各类信息的权重。还有，应该建立对名人网页相关度评价算法的客观评测标准。

整个系统的扩展还准备在多方面进行，比如实体信息库的动态更新、网页库的累增、名人库的扩大等，项目组目前正在实施著名机构/企业的网页检索服务，另外，相关度评价算法还可以用于其它的网络检索服务，比如数字图书馆系统等。

参 考 文 献

[1] Douthat. The Message Understanding Conference Scoring Software User's Manual. MUC-7 Proceedings. SAIC 1999

[2] Jiawei Han, Micheline Kamber. Data Mining—Concept and Tachniques. Academic Press, 2000

[3] Ray, Deborah S. Mastering Html 4.0 1998

[4] 孙斌： 中文信息提取系统设计与若干相关基础问题的研究 博士后研究工作报告 2002.5

[5] 孙斌： 信息提取技术概述 术语标准化与信息技术，2002.3，2002.4，2003.1

[6] 谷红英，俞士汶： CCD 及其应用 广西师范大学学报(p98-103) 2003.1