



AskTheWeb

我们对于问题回答系统的一次实验

北京大学信息科学学院

作者：许丞，彭瀚，李双峰，马龙

2004年11月



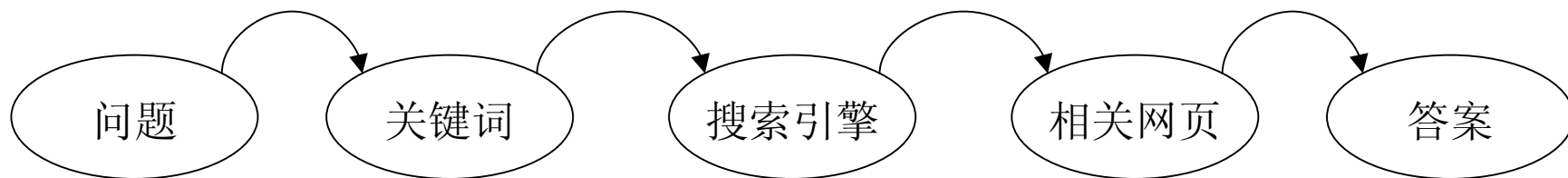
提纲

1. 为什么要QA?
2. 互联网上的QA系统
3. AskTheWeb: 我们的特点
4. 系统结构
5. 测试结果
6. 待完善的地方和将来的工作

为什么要QA：搜索引擎的不足

■ 使用全文检索的搜索引擎：

- 用户输入的关键词是决定能否找到目标网页的决定因素，由问题转换到查询关键词，既不方便，还需要有构思恰当的关键词的经验
- 搜索引擎返回大量网页，用户在其中查找相关信息需要花费很多精力





为什么要QA：搜索引擎的不足

- 具有分类目录的搜索引擎
 - 对网站的简短描述无法提供足够的信息：我在这个站点上能不能找到答案？
 - 网站分类过程耗费大量人力



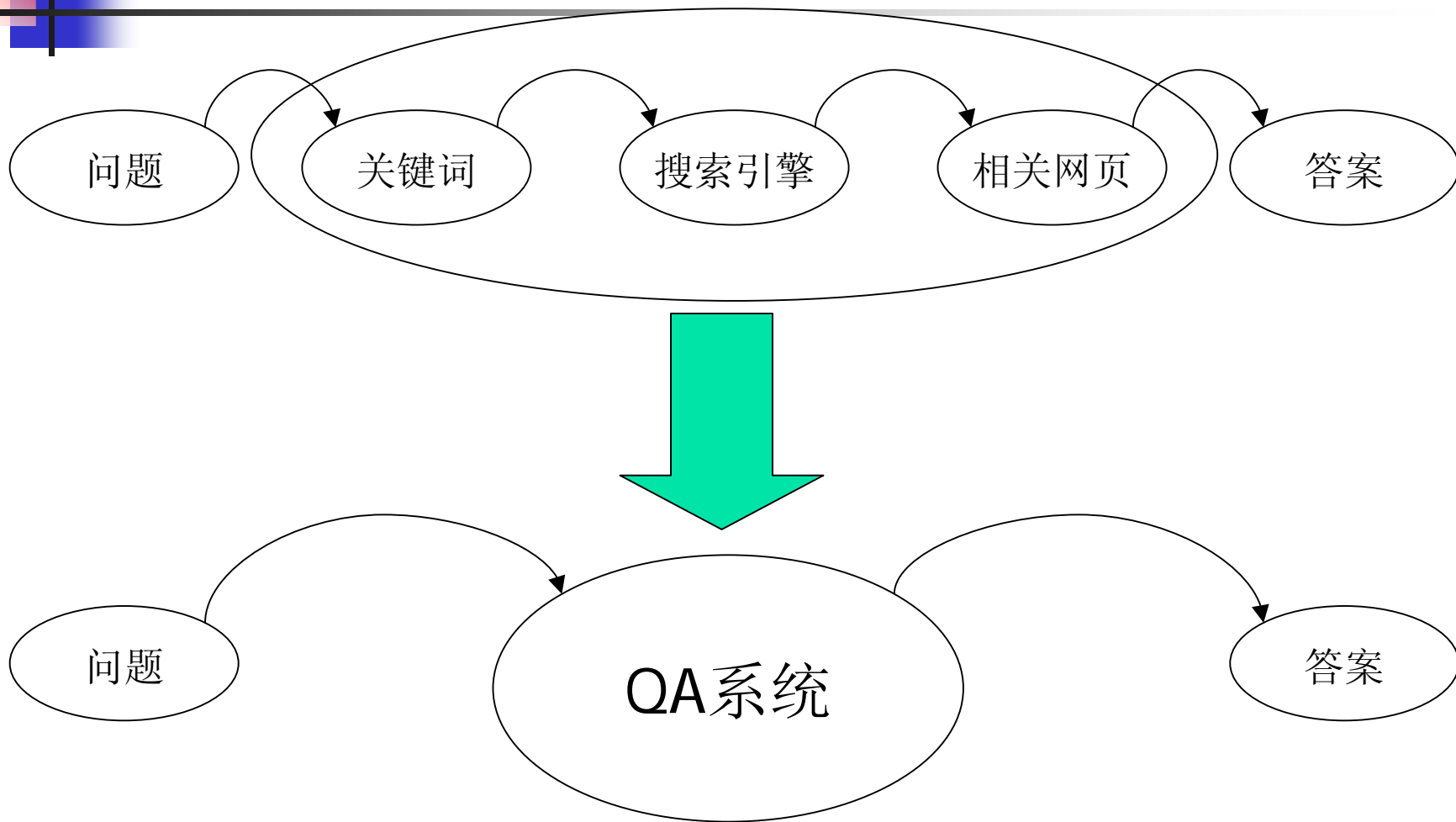
为什么要QA：搜索引擎的不足

- 中文Yahoo!在“科学>地理”下提供的站点目录：

简体中文 GB:

- [大峡谷考察](#) 中 - 介绍中国的雅鲁藏布大峡谷的地理、风情、人文、自然遗产等。
- [地理科普](#) 中 - 提供地理新闻、人文地理、区域地理、地理辞典、地理人物等。
- [地理天空](#) 中 - 含自然地理、人文地理、乡土地理、地理图片、以及论文等。
- [世界地理频道](#) 中 - 介绍地球知识、地球生物、七大洲等信息。
- [信息共享政策机制与管理办法研究](#) 中 - 从事地理信息共享研究，含课题介绍、论文汇编、人才培养、期刊文章、研究成果。
- [中国地理信息产业资讯网/中国测绘信息网](#) 中 - 含3S论坛、SDI和数字地球、行业论坛、地理信息应用、测绘文摘数据库等。
- [地理家园](#) - 含时事地理、网络课例、资源中心、地理。
- [地理频道工作室](#) - 含地理论坛、地理新闻、地理教学、资源中心。
- [国家地理](#) - 含国家地理报导、走遍中国、民风民俗、自然探索。

为什么要QA: QA的理想





为什么要QA: QA的理想

- 用自然语言提问，而不需要考虑关键字和关键字的组合
- 系统自动从相关网页中提取答案，而不需要用户在数十个页面中查找

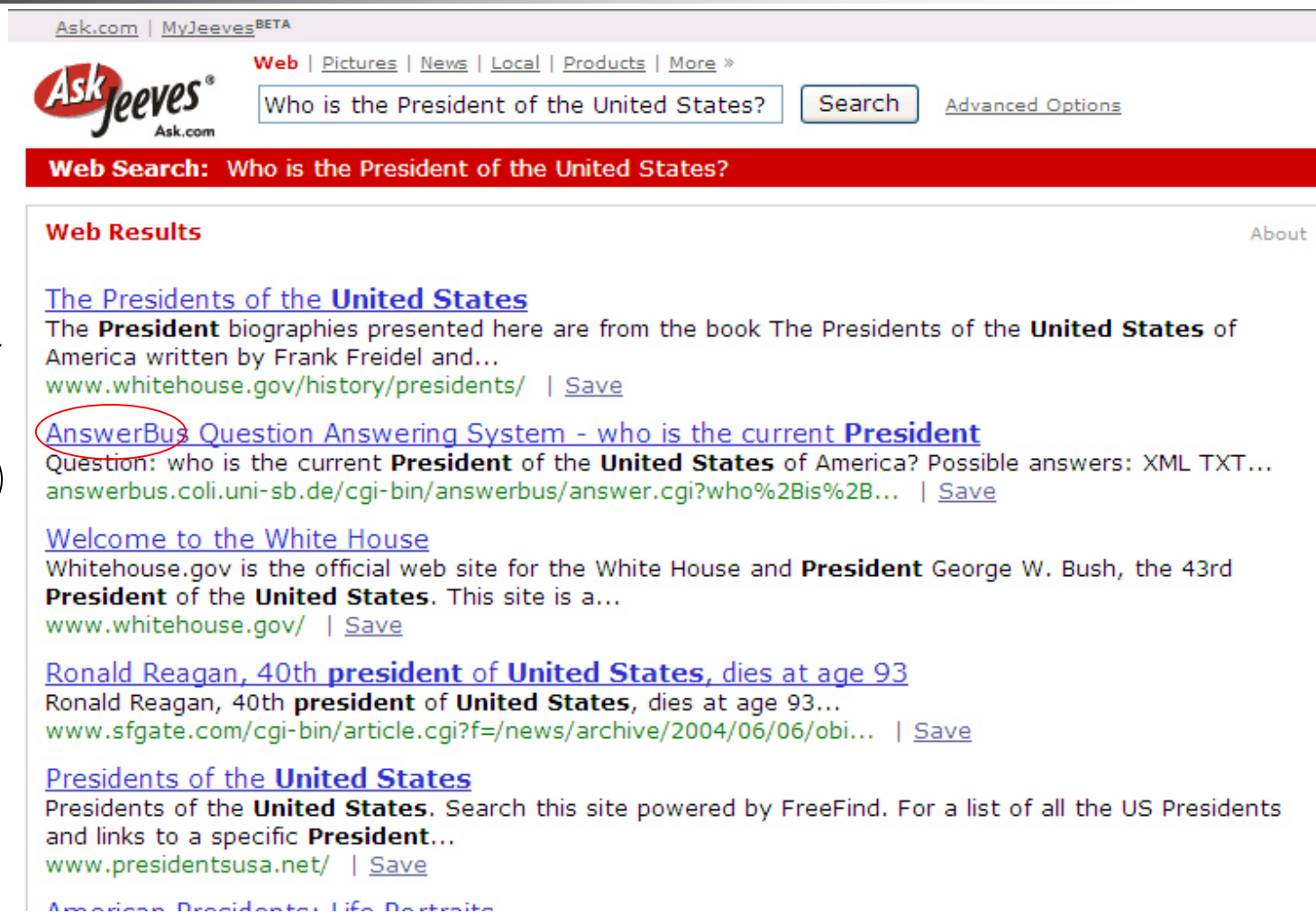


互联网上的QA系统：现实

- 限于目前计算机的智能水平，QA系统只能接受一些基于事实的、短答案的问题。
 - 无需推理即可获得答案
 - 具有确定的答案
 - 对某个属性的简短回答，而不是对过程的描述

互联网上的QA系统：AskJeeves

■ www.ask.com



Ask.com | MyJeeves^{BETA}

Ask Jeeves[®]
Ask.com

Web | Pictures | News | Local | Products | More »

Who is the President of the United States? Search Advanced Options

Web Search: Who is the President of the United States?

Web Results About

[The Presidents of the United States](#)
The **President** biographies presented here are from the book The Presidents of the **United States** of America written by Frank Freidel and...
www.whitehouse.gov/history/presidents/ | Save

[AnswerBus Question Answering System - who is the current President](#)
Question: who is the current **President** of the **United States** of America? Possible answers: XML TXT...
answerbus.coli.uni-sb.de/cgi-bin/answerbus/answer.cgi?who%2Bis%2B... | Save

[Welcome to the White House](#)
Whitehouse.gov is the official web site for the White House and **President** George W. Bush, the 43rd **President** of the **United States**. This site is a...
www.whitehouse.gov/ | Save

[Ronald Reagan, 40th president of United States, dies at age 93](#)
Ronald Reagan, 40th **president** of **United States**, dies at age 93...
www.sfgate.com/cgi-bin/article.cgi?f=/news/archive/2004/06/06/obi... | Save

[Presidents of the United States](#)
Presidents of the **United States**. Search this site powered by FreeFind. For a list of all the US Presidents and links to a specific **President**...
www.presidentsusa.net/ | Save

American Presidents: Life Portraits

相关网页而不是问题答案



互联网上的QA系统: AnswerBus

AnswerBus

Who is the President of the United States?

Ask

Type in your question in English, French, Spanish, German, Italian or Portuguese.

Question:

Who is the President of the United States?

www.answerbus.com, 给出可能含有答案的句子

Possible answers: [XML](#) [TXT](#)

- [The agency symbol assigned to the President of the United States is PR followed by the number corresponding to the ordinal number of succession to the presidency as PR 42, Bill Clinton, 42nd president of the United States.](#)
- [Background and Purpose New York City is often visited by the President and Vice President of the United States, as well as visiting heads of foreign states or foreign governments, on the average of 12 times per year.](#)
- [\(quarterly\) Executive Office of the President / Vice President of the United States](#)
- [This zone extends bank to bank while the President of the United States addresses, or is in attendance at, the United Nations General Assembly.](#)
- [NS 1.61: POWRE: Professional Opportunities for Women in Research and Education \(annual\) President of the United States](#)
- [Now, Therefore, I, Ronald Reagan, President of the United States of America, do hereby proclaim October 27, 1982, as a Day of National Celebration of the one hundred twenty-fifth anniversary of the birth of Theodore Roosevelt.](#)



互联网上的QA系统：MIT START

- <http://www.ai.mit.edu/projects/infolab/ailab.html>
 - 直接给出答案
-

START's reply

==> Who is the President of the United States?

The forty-third president of the United States is [George Walker Bush](#).

Source: [Internet Public Library](#)

- [Go back to the START dialog window.](#)



AskTheWeb: 我们的特点

- 面向中文的QA系统
- 利用网页的冗余信息提取答案
- 利用搜索引擎和其他异构信息源
- 答案类型猜测和概念匹配
- 给出基于短语的答案，而不是句子
- 从一个课程设计发展而来

AskTheWeb界面

Ask The Web!

请输入问题 非洲最高峰是什么?

提问

答案列表

答案	频次
乞力马扎罗山	15
罗峰	8
非洲乞力马扎罗山	5
厄尔布鲁士峰	5
罗山	4
珠穆朗玛峰	4
四川	3
亚峰	3
北美麦金利峰	2
亚洲珠穆朗玛峰	1

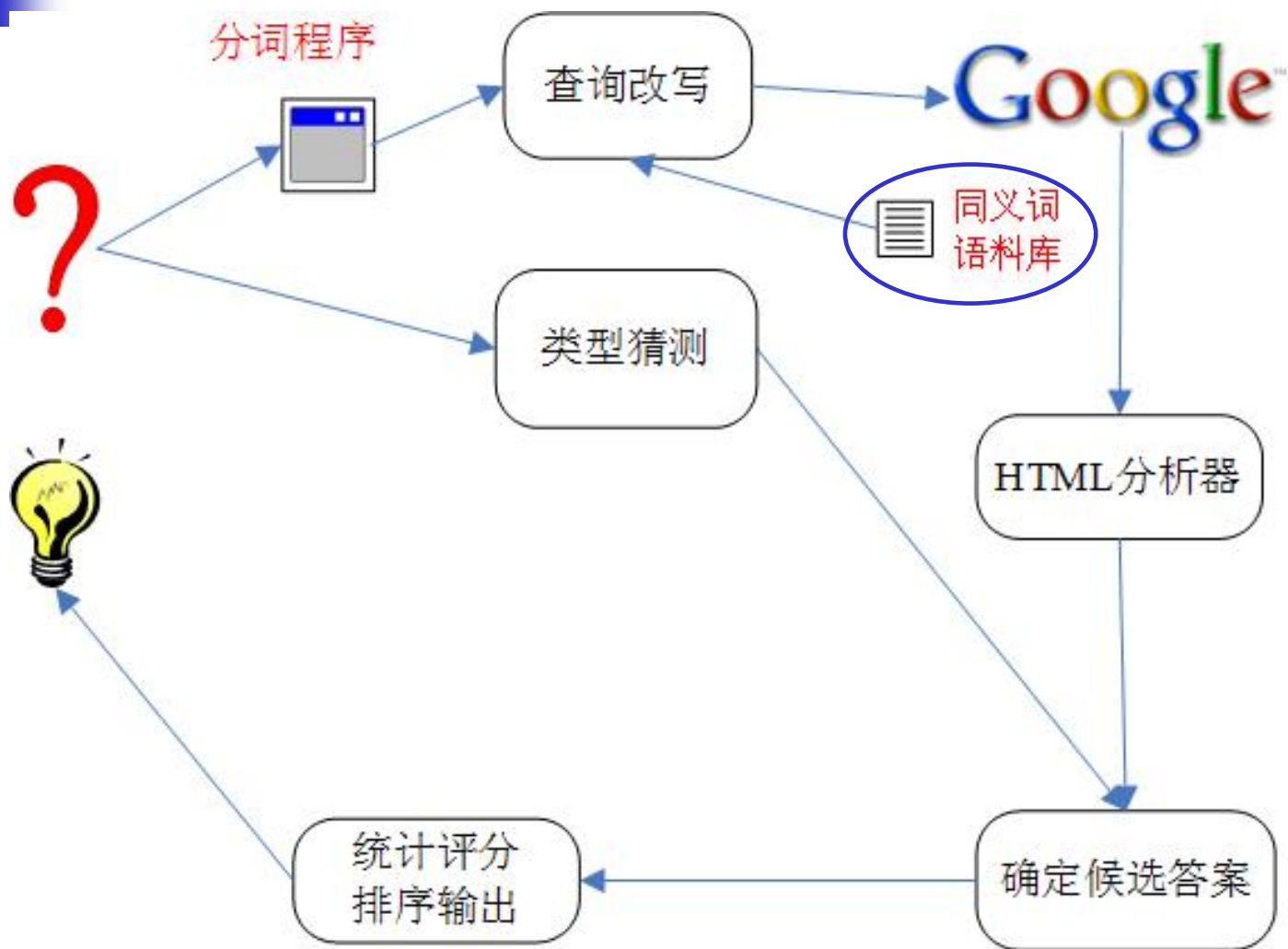
人名

- 北京大学校长是谁?
- AC米兰队长是谁?
- 武汉市市长是谁?
- “大江东去浪淘尽”是谁写的?
- 谁写了《桃花扇》?
- 《追随智慧》的作者是谁?
- 《地下地》的导演是谁?

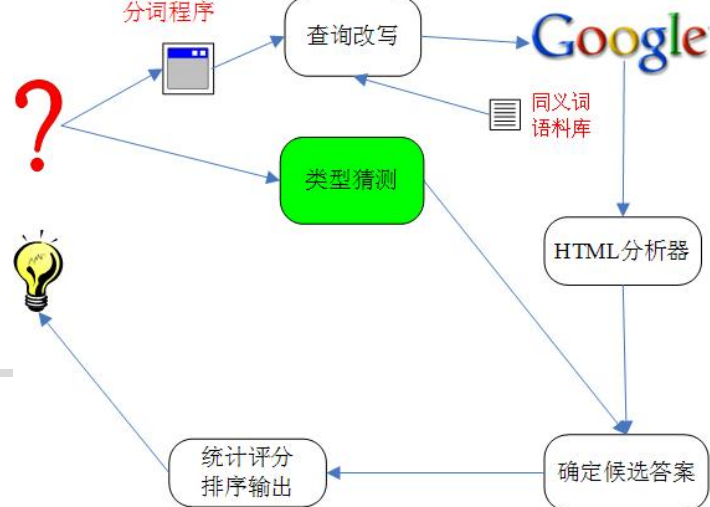
地理知识: 山峰、河流、方位、高度、长度

- 世界第一高峰是哪一座?
- 非洲最高峰是什么?
- 中国最长的河是什么河?
- 中南民族学院在哪个地方?
- “乞力马扎罗山”有多高?
- “珠穆朗玛峰”有多高?
- 长江全长有多长?

AskTheWeb系统结构



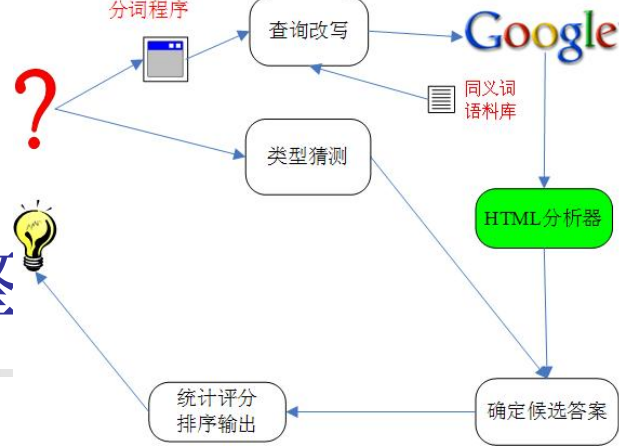
系统结构：类型猜测



- 将查询问句分词，提取关键词，猜测用户的问句的类型(问是谁，问什么地方...)及答案类型(数字，地名，人名，物品...)
 - 原型系统中采用的方法：正则表达式匹配
 - “...是谁”→问人名
 - “...是哪一座”→问山峰名

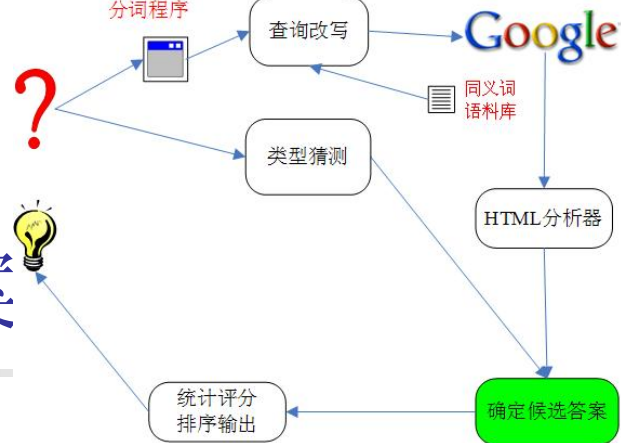
更好的方法 = ?

系统结构：查询搜索引擎



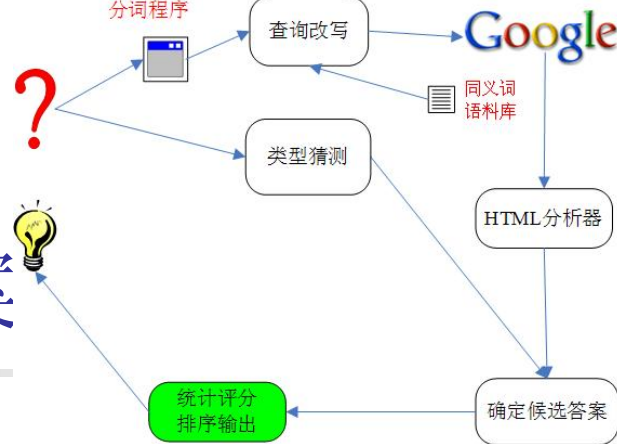
- 向Google或其他搜索引擎发出查询
 - 不同的查询条件应该有不同的weight
- 获取Google返回的查询结果页面，得到相关网页的摘要
 - 因为效率的原因，只分析Google返回页面上的网页摘要(称为summary)而不再获取原始网页
 - 因为网络的限制，这一步中获得的网页通常不超过10张，最多100个摘要

系统结构：确定候选答案

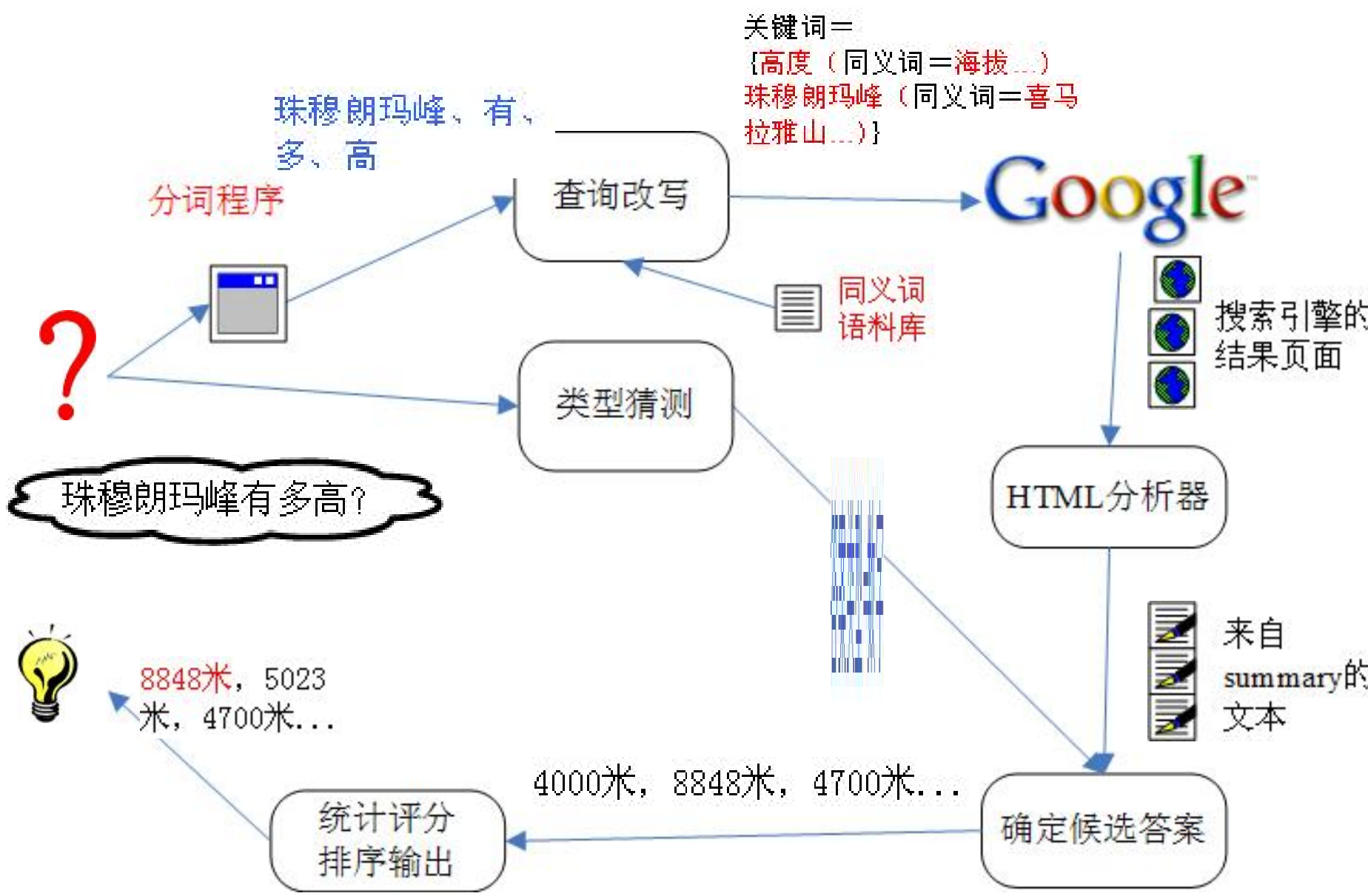


- 首先对summary进行中文切词，并标注词性
- 切词程序已经提供了粗略的类型信息
 - 例如：人名，地名，数字，等等
- 进一步的分类
 - 原型系统中采用的方法：正则表达式匹配
 - 13312345678 -> 13d9 -> 国内移动电话号码
 - 珠穆朗玛峰 -> *峰 -> 山峰名
 - 更好的方法：WordNet

系统结构：统计候选答案



- 将与猜测的答案类型相关的词进行统计，并根据提交查询的关键词的权对结果进行打分
 - 词频统计
 - 带权的词频统计
- 将得分最高的词作为问题的答案，同时输出得分较低的几个作为参考
 - 在原型系统中，我们为简化起见，只实现对词的出现频率进行统计，不再应用加权
 - 输出所有可能答案的可能性百分比（其实就是出现频率百分比）



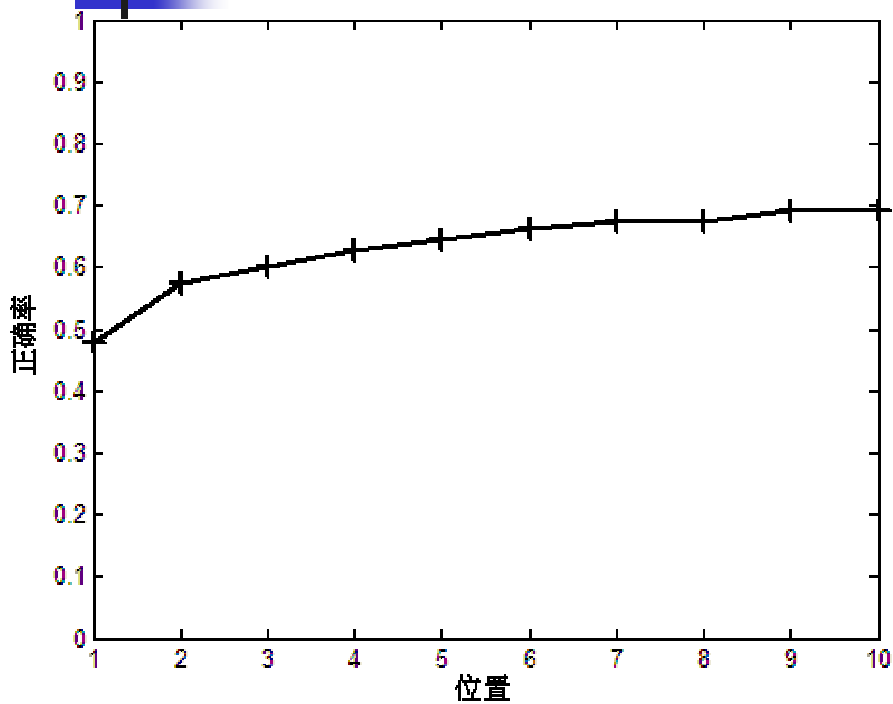


测试结果

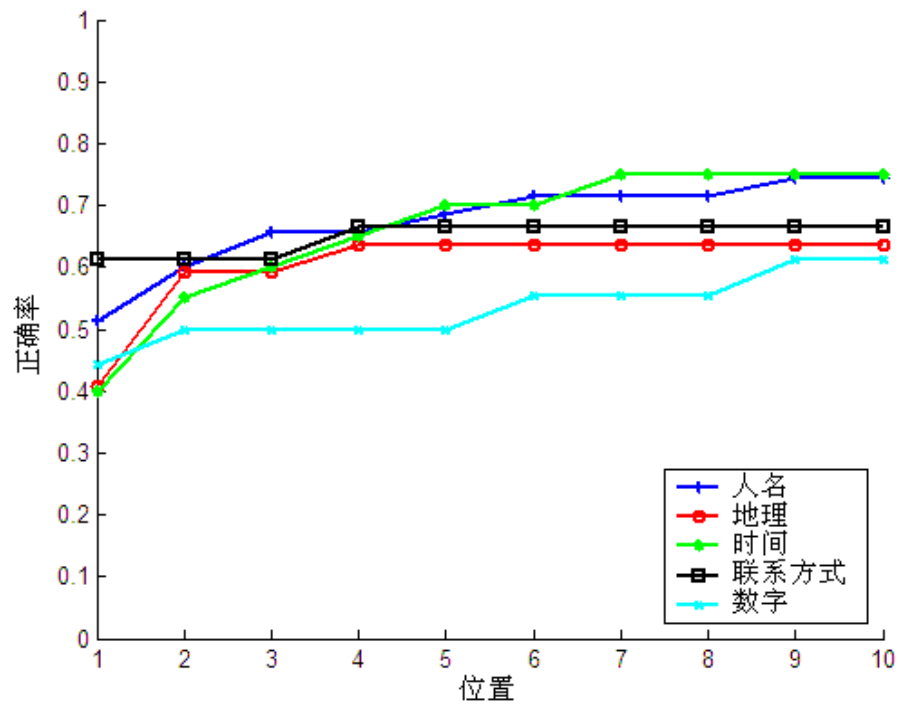
- 人名类问题35个，如：“大江东去浪淘尽”是谁写的？
- 地点类问题22个，如：世界第五高峰是哪一座？
- 时间类问题20个，如：美国独立日是哪一天？
- 通讯方式问题18个，如：北京大学计算机系教务的电话号码是什么？
- 数字类问题18个，如：上海金茂大厦有多高？

答案分别为：苏轼，马卡鲁峰，7月4日，62751890，420米

测试结果



在第x个位置前出现
正确答案的比例



在第x个位置前出现
正确答案的比例
(分类)



总结：

- AskTheWeb: QA领域中的一次尝试
- 原型系统为今后的扩展搭建了平台
 - 前处理：完善的同义词库，WordNet，类型猜测
 - 统计公式、算法
 - 向多搜索引擎和本地数据库的移植



AskTheWeb的缺点和将来的工作

- 缺少有力的理论支持和实验数据
- 基于统计的打分模型，没有建立一个完善的理论模型
- 没有完善的类型猜测子系统
- 缺少足够的同义词库



AskTheWeb

谢谢!