使用关系模式进行半结构化搜索

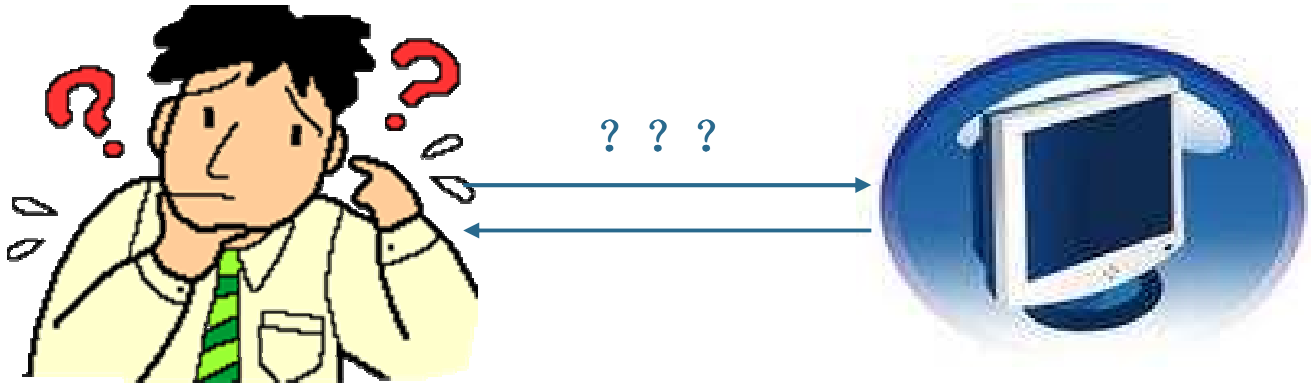
指导教师：赵燕平 张华平

报告人：马静

http://hi.baidu.com/drkevinzhang/

Contents

- 为什么要使用关系模式进行半结构化搜索？



- 为什么要使用关系模式进行半结构化搜索？

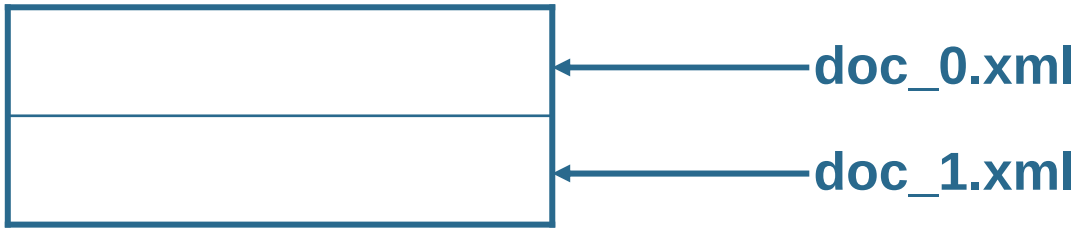
- 用什么样的关系模式进行半结构化搜索？
  - XML（可扩展性标记语言）
    1、跨平台，可移植
    2、简单，易于掌握
    3、可以自定义标签

What

<DOC>
<DOCNO> WSJ870323-0180 </DOCNO>
<HL> Italy's Commercial Vehicle Sales </HL>
<DD> 03/23/87 </DD>
<DATELINE> YURIN,Italy </DATELINE>
<TEXT>
Commercial-vehicle sales in Italy rose I 1.4 % in February from a year earlier,to 8,848 units,according to provisional figures from the Italian Association of Auto Makers.
</TEXT>
</DOC>

- 用什么样的关系模式进行半结构化搜索？
  - 支持任意xml模式的静态关系模式

1、静态关系存储模式负责存储所有不同的xml路径，其中每篇文档对应关系中相互独立的行。

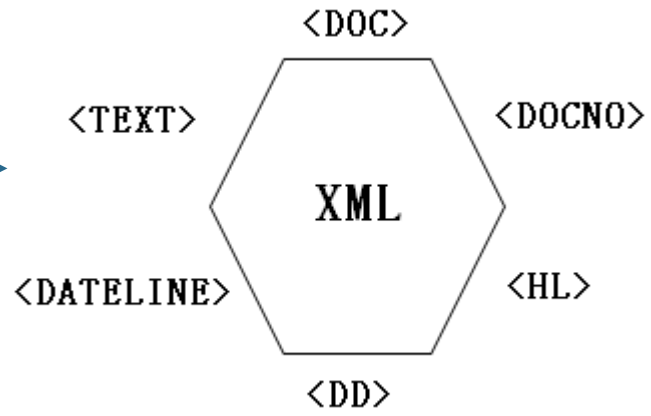


❖ 用什么样的关系模式进行半结构化搜索？

- 支持任意xml模式的静态关系模式

2、每一行对应xml结构图中的一条边，这种静态关系模式可以存储任意的xml文档。

tagId	tag
10	DOC
11	DOCNO
12	HL
13	DD
14	DATELINE
15	TEXT



- 用什么样的关系模式进行半结构化搜索？
  - 支持任意xml模式的静态关系模式

    3、Xml文档层次结构的保存方式十分变通，在每篇文档索引至数据库时，仅需表中的信息便可重建。

❖怎样使用关系模式进行半结构化搜索？

<DOC>
<DOCNO> WSJ870323-0180 </DOCNO>
<HL> Italy's Commercial Vehicle Sales </HL>
<DD> 03/23/87 </DD>
<DATELINE> YURIN,Italy </DATELINE>
<TEXT>
Commercial-vehicle sales in Italy rose I 1.4 % in February from a year earlier,to 8,848 units,according to provisional figures from the Italian Association of Auto Makers.
</TEXT>
</DOC>

怎样使用关系模式进行半结构化搜索？

1、存储xml元数据

TAG_NAME

tagId	tag
10	DOC
11	DOCNO
12	HL
13	DD
14	DATELINE
15	TEXT

存储xml集合中所有唯一标签的命名

TAG_PATH

tagId	path
10	[DOC]
11	[DOC,DOCNO]
12	[DOC,HL]
13	[DOC,DD]
14	[DOC,DATELINE]
15	[DOC,TEXT]

存储在xml文档中出现的不同访问路径

ATTRIBUTE

AttrId	attribute
7	LAUNGUAGE

存储所有属性的命名

- 怎样使用关系模式进行半结构化搜索？

2、追踪xml文档

每当为新文件建立索引时，该文件就得到一个唯一的标识符作为固定值，每个值对应一个xml文件。

DOCUMENT

docId	fileName
6	doc_0.xml
7	doc_1.xml

- 怎样使用关系模式进行半结构化搜索？

3、建立xml索引

存储当前标签的父标签id

访问路径是否由元素或属性所终结

INDEX

记录标签的出现顺序

最底层标签内的值

Id	parent	path	type	tagId	AttrId	docId	pos	value
41	0	10	E	10	1	6	0	NULL
42	41	11	E	11	1	6	0	WSJ870323-0180
43	41	12	E	12	1	6	0	Italy's commercial…
44	41	13	E	13	1	6	0	03/23/87
45	41	14	E	14	1	6	0	TURIN,Italy
46	41	15	E	15	1	6	0	Commercial-vehicle…
47	46	15	A	15	7	6	0	English

❖怎样使用关系模式进行半结构化搜索？

<DOC>
<DOCNO> WSJ870323-0180 </DOCNO>
<HL> Italy's Commercial Vehicle Sales </HL>
<DD> 03/23/87 </DD>
<DATELINE> YURIN,Italy </DATELINE>
<TEXT>
Commercial-vehicle sales in Italy rose I 1.4 % in February from a year earlier,to 8,848 units,according to provisional figures from the Italian Association of Auto Makers.
</TEXT>
</DOC>

INDEX

Id	parent	path	type	tagId	AttrId	docId	pos	value
41	0	10	E	10	1	6	0	NULL
42	41	11	E	11	1	6	0	WSJ870323-0180
43	41	12	E	12	1	6	0	Italy's commercial…
44	41	13	E	13	1	6	0	03/23/87
45	41	14	E	14	1	6	0	TURIN,Italy
46	41	15	E	15	1	6	0	Commercial-vehicle…
47	46	15	A	15	7	6	0	English

- 怎样使用关系模式进行半结构化搜索？

TAG_NAME

tagId	tag
10	DOC
11	DOCNO
12	HL
13	DD
14	DATELINE
15	TEXT

TAG_PATH

tagId	path
10	[DOC]
11	[DOC,DOCNO]
12	[DOC,HL]
13	[DOC,DD]
14	[DOC,DATELINE]
15	[DOC,TEXT]

- 怎样使用关系模式进行半结构化搜索？

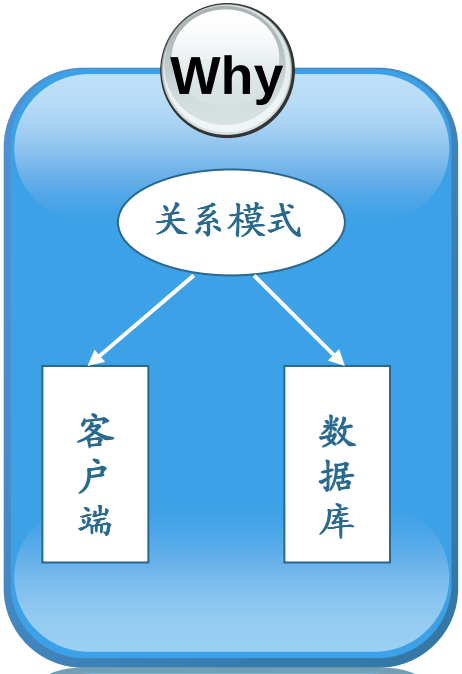
DOCUMENT

docId	fileName
6	doc_0.xml
7	doc_1.xml

❖怎样使用关系模式进行半结构化搜索？

<DOC>
<DOCNO> WSJ870323-0180 </DOCNO>
<HL> Italy's Commercial Vehicle Sales </HL>
<DD> 03/23/87 </DD>
<DATELINE> YURIN,Italy </DATELINE>
<TEXT>
Commercial-vehicle sales in Italy rose I 1.4 % in February from a year earlier,to 8,848 units,according to provisional figures from the Italian Association of Auto Makers.
</TEXT>
</DOC>

Summary



What

1. xml的优点:
   跨平台
   简单易懂
   自定义标签
2. 支持所有xml模式的静态关系模式

How

1. 存储xml元数据
2. 跟踪xml文档
3. 建立xml索引

Thank You !

Contact
Email: kevinzhang@bit.edu.cn

Welcome to visit my blog

http://hi.baidu.com/drkevinzhang/