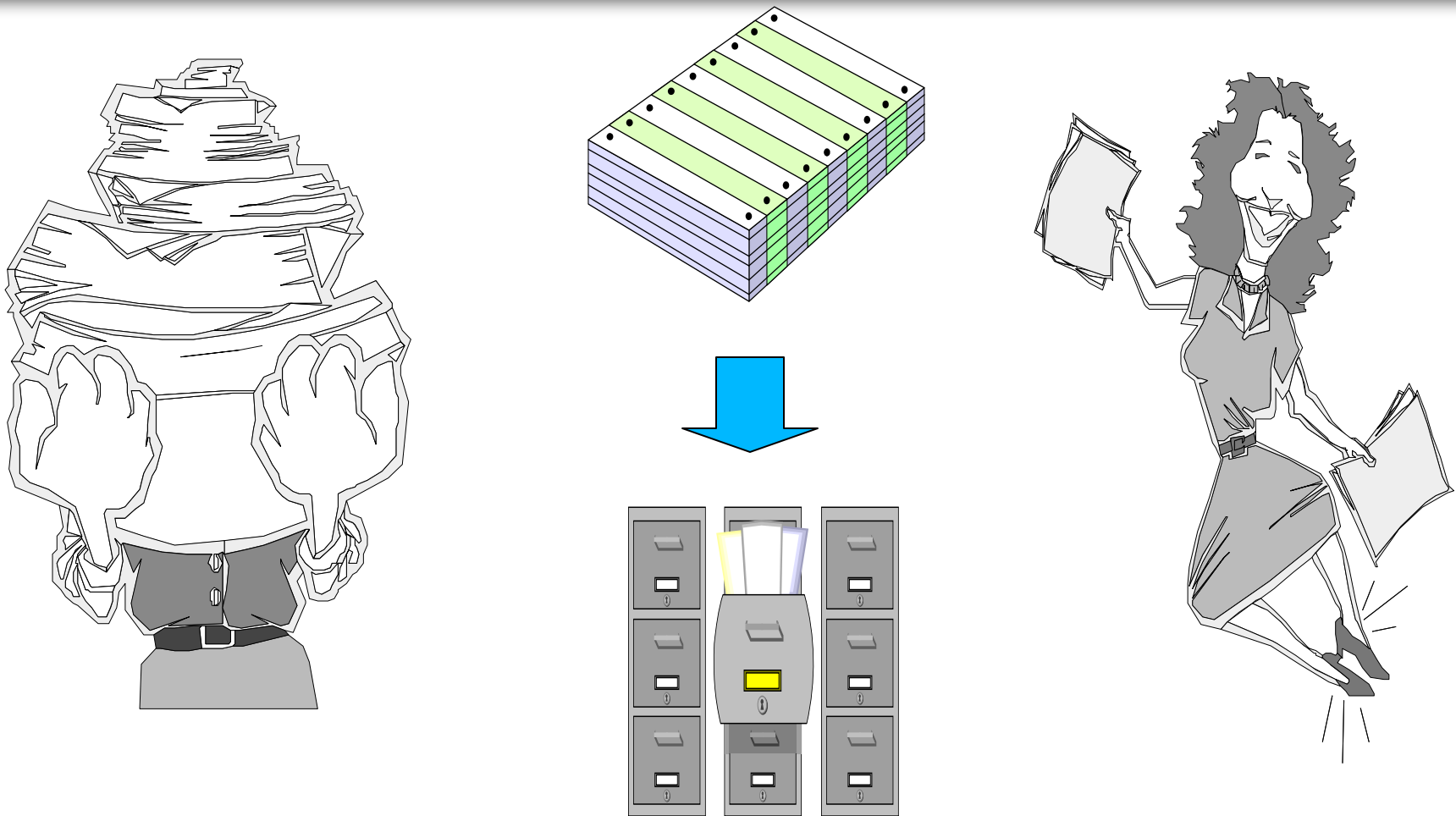


Towards Knowledge Extraction from Weblogs and Rule-based Semantic Querying

Xi Bai, Jigui Sun, Haiyan Che, Jin Wang
College of Computer Science and Technology
Jilin University

Email: xibai@email.jlu.edu.cn

Towards Knowledge Extraction from Weblogs and Rule-based Semantic Querying



Towards Knowledge Extraction from Weblogs and Rule-based Semantic Querying

- Goal:
 - ◆ Propose a novel knowledge extraction framework for mining bolgs.
 - ◆ Implement a prototype which can automatically extract the machine-understandable knowledge after the pre-clustering of blog pages.
- Motivation:
 - ◆ More and more companies and organizations begin to put their information on Web logs (blogs) nowadays.

Towards Knowledge Extraction from Weblogs and Rule-based Semantic Querying

- ◆ Since the data extracted by traditional IE techniques are usually not machine-understandable, it can not be processed by computers and remains raw and redundant
- Challenge:
 - ◆ How to crawl blog document efficiently
 - ◆ How to extract the Information Content Blocks of Interests(ICBI)
 - ◆ How to let the users take part in the reasoning

Overview

- Background:
 - ◆ Snippet clustering
 - ◆ Web ontology language (OWL)
 - ◆ Reasoning
 - ◆ Regular Expression
- Experiments
 - ◆ Extract information from the retrieved blogs without domain ontologies and with domain ontologies respectively

Overview

- Related work
- Future work
- Conclusions

Snippet Clustering

- Crawling
 - ◆ Google AJAX Search APIs[1]
- Features Extraction
 - ◆ Vector Space Model (VSM)
 $(tf_{w_1} \cdot idf_{w_1}, tf_{w_2} \cdot idf_{w_2}, \dots, tf_{w_n} \cdot idf_{w_n})$
 - ◆ Suffix Array
 - The suffix array of a string is the array of all its suffixes sorted by lexicographically

[1] <http://code.google.com/apis/ajaxsearch>

Snippet Clustering

- Dimensions Reduction
 - ◆ The dimension of feature vectors extracted from Chinese documents is usually too large (maybe larger than 50000)
 - ◆ Latent Semantic Indexing (LSI)

$$M_k = U_k S_k V_k^T$$

- ◆ Singular Value Decomposition (SVD)

$$d^* = d^T V_k^T S_k^{-1}$$

Snippet Clustering

- Category-Names Extraction

- ◆ Cosine distance

$$D = U_k^T R$$

- Snippets Distribution

$$C = P^T R$$

- Interesting Block Extraction

Snippet Clustering

Definition 1. *A page is judged to be a blog iff a sequence of entries that are articles for a day can be extracted from the page.*

- Regular Expression

- ◆ “22-May-2007” “22 May 2007”
- ◆ $(\backslash d+)\backslash .?-?\backslash s?([\backslash d\backslash .\backslash s-]+)\backslash .?-?\backslash s?(\\d\{4\})$

Definition 2. *String \mathcal{E} is a snippet and string \mathcal{B} is an information block. Function $ws(s)$ can return the set containing all the words appearing in string s . \mathcal{E} is contained by \mathcal{B} iff $\text{coverage}(\mathcal{E}, \mathcal{B}) \geq \rho$, where $\text{coverage}(\mathcal{E}, \mathcal{B}) = \frac{|ws(\mathcal{E}) \cap ws(\mathcal{B})|}{|ws(\mathcal{E})|}$.*

Knowledge Generation

- $K=(subject, predicate, object)$
- Word segmentation and Part of Speech tagging
 - ◆ ICTCLSAS
- Identify the subject, the predicate (core verb), and the object for each sentence
 - ◆ ICTPROP

Knowledge Generation

- Sentence Types:

1. “做(Do)” statements (ex. *CPCC promulgated a bulletin yesterday.*);
2. “是(Be)” statements (ex. *The ID number of CPCC stock is 600028.*);
3. “成为(Become)” statements (ex. *The price of CPCC becomes 12.33.*).

- Name Mapping

- ◆ Establish a Name Dictionary based on HowNet[2]

[2] www.keenage.com

Example of Name Mapping

```
<?xml version="1.0" encoding="UTF-8" ?>
<INFORMATION>
  <CORPERATION>
    <法定名称>安徽皖通高速公路股份有限公司</法定名称>
    <英文名称>Anhui Expressway Company Limited</英文名称>
    <成立日期>1996-08-15</成立日期>
    <上市市场>上海证券交易所</上市市场>
    <上市日期>2003-01-07</上市日期>
    <所属行业>交通运输、仓储业</所属行业>
    <注册资本>165,861.00</注册资本>
    <同行业公司数>93</同行业公司数>
    <法人代表>王水</法人代表>
    <职工总数>1081</职工总数>
    <董事会秘书>谢新宇</董事会秘书>
    <公司电话>0551-5338697</公司电话>
    <董秘电话>0551-5338681</董秘电话>
    <公司传真>0551-5338696</公司传真>
    <董秘传真>0551-5338696</董秘传真>
    <公司电子邮件>wtgs@anhui-expressway.com.cn</公司电子邮件>
    <董秘电子邮件>wtgs@anhui-expressway.com.cn</董秘电子邮件>
    <公司网址>http://www.anhui-expressway.com.cn/</公司网址>
    <报告放置地点>公司本部</报告放置地点>
    <信息披露网址>http://www.sse.com.cn/</信息披露网址>
    <注册地址>安徽省合肥市市长江西路660号</注册地址>
    <境内会计师事务所>普华永道中天会计师事务所有限公司</境内会计师事务所>
    <境外会计师事务所>
    <经营范围>高等级公路建设、设计、监理、收费、养护、管理、技术咨询及广告、
    配套服务、施救、公路运输、仓储、汽车、机械配件、设备维修、建筑装潢、建筑
    材料销售、高新技术产品研制、开发及生产销售主营:持有、经营及开发安徽省境内
    外收费高速公路及公路。</经营范围>
    <公司沿革>安徽皖通是由安徽省高速公路总公司作为独家发起人,经国家体制改革
    委员会体改生[1996]112号文批准,以截止1996年4月30日评估确认的与合宁高
    速相关资产作为出资发起设立的股份有限公司。本公司于1996年8月15日在安徽省
    工商行政管理局登记注册成立。</公司沿革>
  </CORPERATION>
</INFORMATION>
```

```
<?xml version="1.0" encoding="UTF-8" ?>
<INFORMATION>
  <CORPERATION>
    <公司名称>...</公司名称>
    <英文名称>...</英文名称>
    <成立日期>...</成立日期>
    <上市地>...</上市地>
    <上市时间>...</上市时间>
    <所属板块>...</所属板块>
    <注册资本>...</注册资本>
    <同行业公司数>...</同行业公司数>
    <法人代表>...</法人代表>
    <职工总数>...</职工总数>
    <董事会秘书>...</董事会秘书>
    <公司电话>...</公司电话>
    <董秘电话>...</董秘电话>
    <公司传真>...</公司传真>
    <董秘传真>...</董秘传真>
    <公司电子邮件>...</公司电子邮件>
    <董秘电子邮件>...</董秘电子邮件>
    <公司网址>...</公司网址>
    <报告放置地点>...</报告放置地点>
    <信息披露网址>...</信息披露网址>
    <注册地>...</注册地>
    <被审计>...</被审计>
    <被审计>
    <经营范围>...</经营范围>
    <经营状况>...</经营状况>
  </CORPERATION>
</INFORMATION>
```

Rules Construction

- Common Sense Rules:

```
(true, [is_a, s, fine_stock]) :-  
    (true, [is_a, s, stock]), (true, [is_a, c, company]), (true, [is_a, r, report]),  
    (true, [is_a, p, retained_profits]), (true, [issues, c, s]), (true, [release, c, r]),  
    (true, [states, r, p]), (true, [has_value, p, $1]), (true, [>, $1, 0.1])
```

- Personalized Rules:

```
(true, [is_a, s, trouble_stock]) :-  
    (true, [is_a, s, stock]), (true, [current_price, s, $1]),  
    (true, [price_fall_rate, s, $2]), (true, [>, $1, 8]), (false, [<, $2, 0.1])
```

- Knowledge Reasoning

- ◆ KAON2

Search and Results Presentation

- We use the restrained NLP technique to deal with the users' queries. *Restrained* means users use text boxes to generate queries instead of using the natural language
- Each query is composed of two sets:
 - ◆ Condition set
 - ◆ Variable set

Query Construction Algorithm

Algorithm 1 Query Construction Algorithm.

Input: A 3-tuples T containing the users' inputs in 3 boxes.

Output: The condition set and the variable set.

1. Create a 3-tuples $spo = (part_1, part_2, part_3)$;
 2. Create a condition set $condition$ and a variable set $variable$;
 3. Map T to a new 3-tuples $S = (str_1, str_2, str_3)$ based on ND ;
 4. **for** (each string str in S) {
 - 4.1. **if** (str_i is *null*) {
 - 4.1.1. Create variable V_i ; 4.1.2. Assign V_i to $part_i$; 4.1.3. $variable.add(V_i)$;
 - 4.2. **else if** (str_i is the name of a class) {
 - 4.2.1. Get the class C corresponding to str_i ;
 - 4.2.2. Create ins as an instance of C ; 4.2.3. Assign ins to $part_i$;
 - 4.2.4. Create a new literal $L = (\mathbf{true}, [is_a, ins, c])$; 4.2.5. $condition.add(L)$;
 - 4.3. **else if** (str_i is the name of a data type) {
 - 4.3.1. Get the data type dt corresponding to str_i ; 4.3.2. Assign dt to $part_i$;
 - 4.4. **else if** (str_i is the name of a property) {
 - 4.4.1. Get the property pro corresponding to str_i ; 4.4.2. Assign pro to $part_i$;
5. Create a new literal $L = (\mathbf{true}, [part_2, part_1, part_3])$;
6. $condition.add(L)$;
-

Knowledge Generation and Merging

- Knowledge Generation

Definition 3 *The knowledge \mathcal{K} is denoted by a 3 tuple $\mathcal{K}=(\text{subject}, \text{predicate}, \text{object})$:*

- subject is the part that identifies the thing that the statement is about;*
- predicate is the part that identifies the property or characteristic of the subject that the statement specifies;*
- object is the part that identifies the value of that property.*

Based on Definition 3, extracting information from Web pages is actually a process of extracting subjects, predicates and objects

Knowledge Generation and Merging

- Knowledge Generation

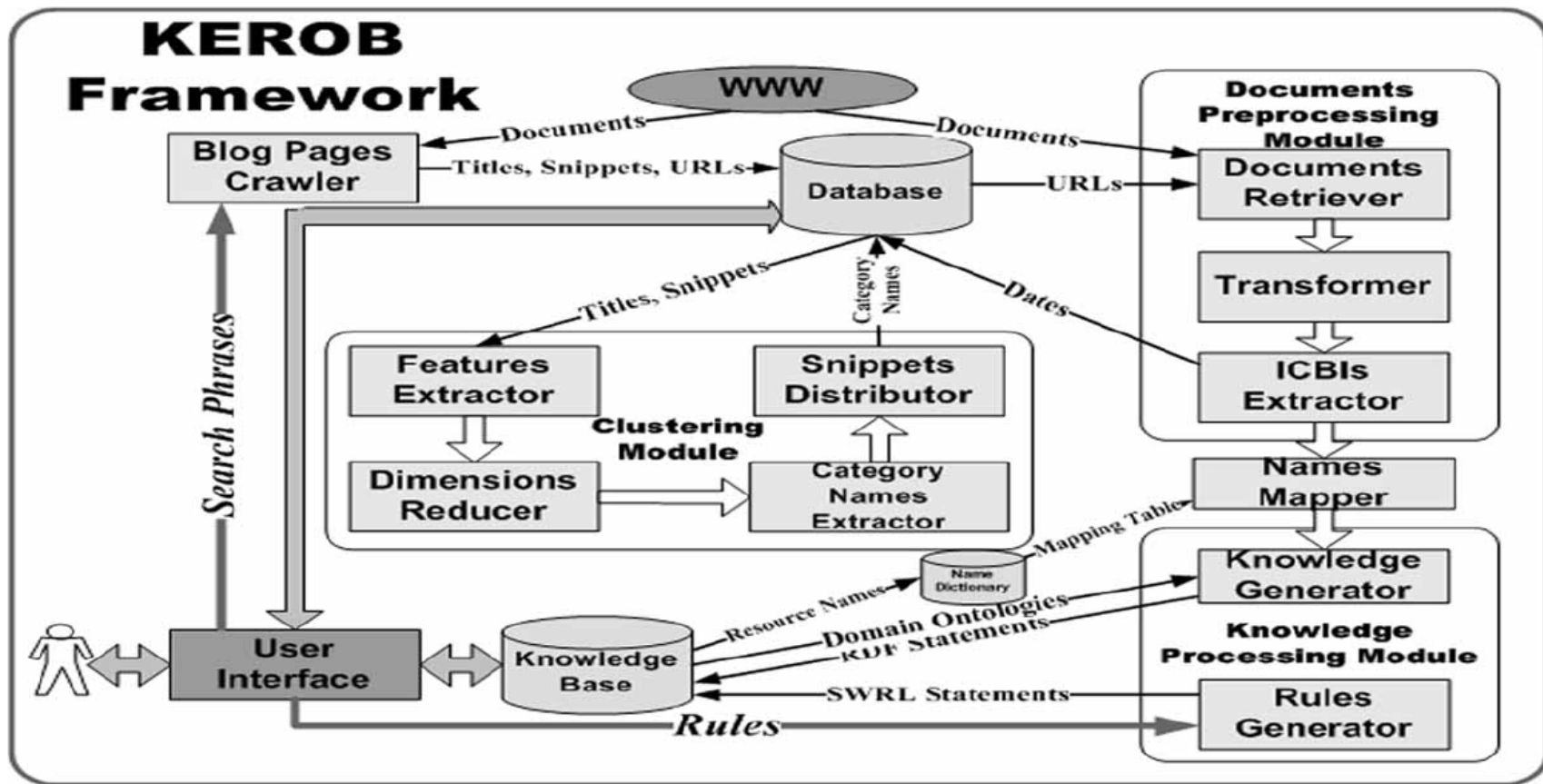
Algorithm 2 Knowledge generation algorithm.

Input:The properties pro appearing in the mapped XML files and the domain ontologies ont .

Output:The RDF files which contain the extracted knowledge.

1. for($i = 0$; $i <$ the number of pro ; $i++$){
 - 1.1. Generate a new individual sub_i ;
 - 1.2. if(pro_i is an object property){
 - 1.2.1 Generate a new individual obj using the class of pro_i 's range;
 - 1.2.2 Add the property pro_i and the value obj to sub_i ;
 - else if(pro_i is a data property){
 - 1.2.3 Add the property pro_i and its value to sub_i ;
 - 1.3. Link the new individual sub_i to the individuals in ont ;
 - 1.4. Merge the new individual sub_i and the individuals in ont ;
2. Return the RDF statements in ont ;
-

Framework for KEROB



Experiments

- Environment
 - ◆ AMD Athlon 64 CPU with 1.9GHz and 1GB of RAM
 - ◆ Java programming language.
- Data
 - ◆ Retrieve 1509 blog documents by invoking Google AJAX Search APIs and taking stock as the inputted search phrase.
 - ◆ Obtain 14 categories (total 1433 documents) except category *other* which contains the unclustered documents (total 76 documents)
 - ◆ Ask 30 users to generate 600 queries for all this categories.

Experiments

- We evaluate the performance of our approach by calculating Recall (R), Precision(P) and F-measure(F).

$$Recall = \frac{|S_{total} \cap |S_{DE}|}{|S_{DE}|} \times 100\%$$

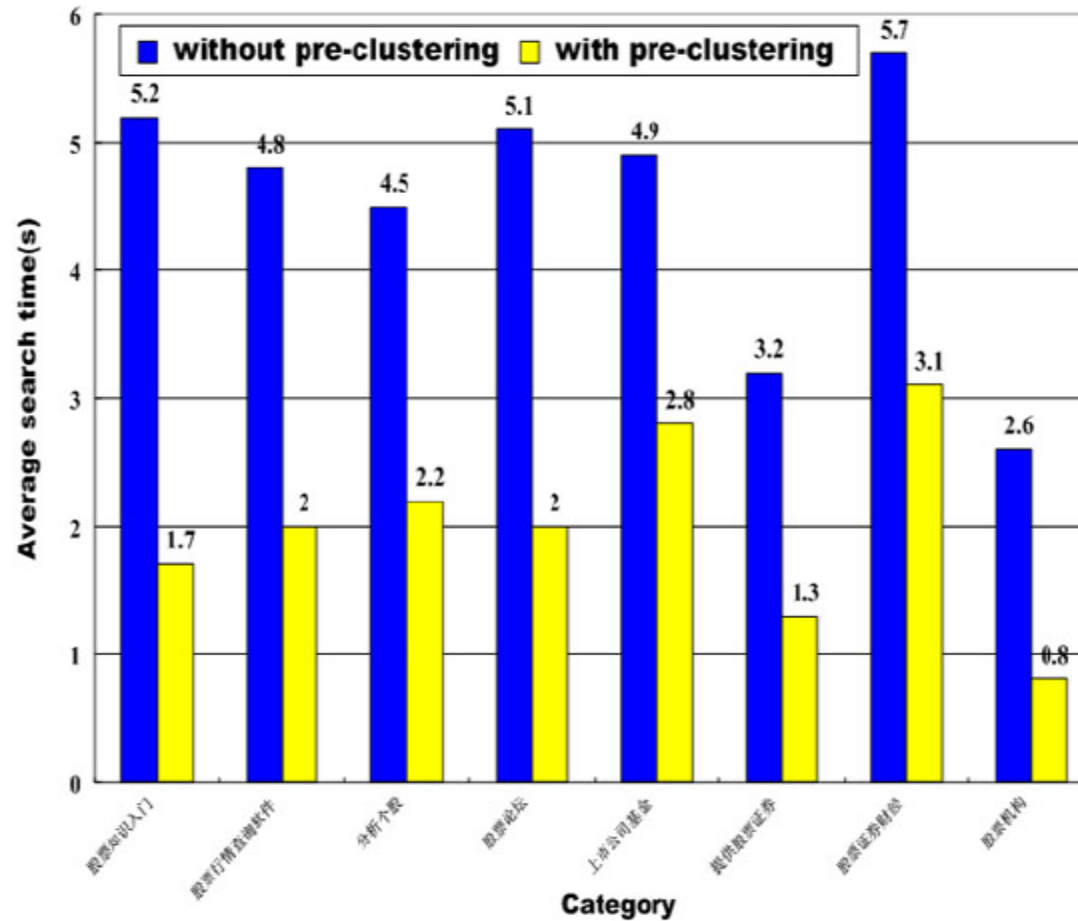
$$Precision = \frac{|S_{total} \cap |S_{DE}|}{|S_{total}|} \times 100\%$$

$$F_{measure} = \frac{2 \times Recall \times Precision}{Recall + Precision} \times 100\% \quad (\beta = 1)$$

Performance

Category Name (In Chinese)	Without Ontologies			With Ontologies		
	Pre.	Rec.	F ₁	Pre.	Rec.	F ₁
知识股票入门	41.16	65.63	53.50	70.35	81.39	75.47
股票行情查询软件	76.25	75.35	75.80	91.45	81.74	86.32
股票网址大全	43.48	62.50	51.28	58.26	70.53	63.81
分析个股	65.05	77.84	70.87	77.08	83.67	80.24
股票论坛	67.87	75.12	71.31	84.94	83.08	84.00
上市公司基金	75.78	81.51	78.54	80.47	84.25	82.32
股票投资	56.82	77.46	65.55	73.55	92.23	81.84
提供股票证券	38.92	59.05	46.92	65.06	76.85	70.46
股票证券财经	60.24	77.77	67.90	72.81	85.28	78.55
股票机构	47.50	70.74	56.84	72.14	81.12	76.37
实时行情	65.06	66.79	65.91	68.03	73.49	70.66
东方财富	65.08	57.75	61.19	73.02	63.01	67.65
股票权证	58.33	66.96	62.35	64.39	71.43	67.73
股票银行	43.93	62.80	51.70	65.32	81.29	72.44

Search Time Comparison

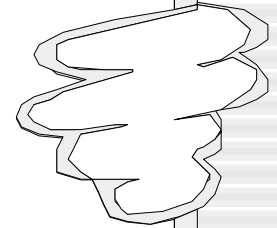


Conclusions

- Future Work
 - ◆ Develop a more general knowledge search engine for blogs by importing other existing domain ontologies
- Conclusion
 - ◆ Our KB is timely updated in the background processing of KEROB. New rules can be created according to the users' specific needs. The experimental results indicate the superiority of our system



**Thank you
&
Questions**



xibai@email.jlu.edu.cn