Ubiquitous Personalized Information Processing & Services

Xindong Wu
Department of Computer Science
University of Vermont, USA
Ubiquitous Personalized Information Processing & Services: Objectives

A positive cycle with
P1: Demand-driven integration of information sources
P2: Mining and analysis
P3: User interest modeling
P4: Security and privacy.
Xindong Wu

Technical Interests: Deduction  Induction
Two Professional Babies
Xindong Wu
TKDE and KDD-07

IEEE TRANSACTIONS ON
KNOWLEDGE AND DATA ENGINEERING

TKDE Editor-in-Chief,
1/1/2005-12/31/2008
Data Mining: Algorithms & Applications
1. Classification


2. Statistical Learning

3. Association Analysis

- **#8. FP-Tree**: Han, J., Pei, J., and Yin, Y. 2000. Mining frequent patterns without candidate generation. In SIGMOD '00.
4. Link Mining


5. Clustering


6. Ensemble Learning

7. Sequential Patterns


- #15. PrefixSpan: J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal and M-C. Hsu. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. In ICDE '01.
8. Integrated Mining

- #16. CBA: Liu, B., Hsu, W. and Ma, Y. M. Integrating classification and association rule mining. KDD-98.
9. Rough Sets

10. Graph Mining

- #18. gSpan: Yan, X. and Han, J. 2002. gSpan: Graph-Based Substructure Pattern Mining. In ICDM '02.
The Top 10 Algorithms

- #1: C4.5
- #2: K-Means
- #3: SVM
- #4: Apriori
- #5: EM
- #6: PageRank
- #7: AdaBoost
- #7: kNN
- #7: Naive Bayes
- #7: Naive Bayes
- #10: CART
Application 1. Diagnosis

- There are 2 alternative hypotheses:
  1. a particular form of cancer (+)
  2. the cancer does not exist (-)
- Prior knowledge over the entire population of people: only 0.008 have this disease
- Lab test is only an imperfect indicator:
  1. A correct positive result in only 98% of the cases when the cancer exists;
  2. A correct negative result with 97% reliability
- To summarize,
  1. P(cancer) = 0.008  P(~cancer) = 0.992
  2. P(+|cancer) = 0.98  P(-|cancer) = 0.02
  3. P(+|~cancer) = 0.03  P(-|~cancer) = 0.97
- Suppose a patient has a positive lab test result. Should we diagnose him/her as having the cancer?

  The answer is no!
Application 2: OIDM
 OIDM (2)

Data Mining Tools

Classification Tools

- Tree Construction Tools
  - C4.5

Rule Generating Tools

- C4.5Rules, OneR, Prism, HCV

Association Analysis Tools

- Apriori

Clustering Tools

- CobWeb
- K-Means
Application 3: User Modeling
Application 4: Noise Handling
The Russell Paradox

- Nobel laureate in Literature 1950
- Also a philosopher, logician and mathematician
- There was once a barber,
  - Wherever he lived, all of the men in his town either shaved themselves or were shaved by the barber. And the barber only shaved the men who did not shave themselves.
- Did the barber shave himself?

- Can we solve the Russell paradox?
- Yes,
  - Mathematically,
    - type theory (by Russell) and axiomatic set theory
  - In data mining, we treat it as systematic noise!
More to Logic …

- Can we solve the Russell Paradox (in data mining)? If so, how?
  - Change the question data before answering
- It was said (by Russell?) everything follows logic except
  - Love
  - Religion
  - Wars
- What is the next step for noise-tolerant data mining?
  - Domain knowledge
  - Noise profiling
  - Unknown noise types.
**Application 5:**

Web News Filtering and Summarization
 Welcome

请输入URL： [ ] [ ]

项目实现成功：新闻网页过滤并提取出具有语义联系的关键语句！

下面是一些新闻页面，可供您试用：

http://www.xp.com.cn/20071225/20001011.htm
http://www.xp.com.cn/20071226/20000600.htm
http://www.xp.com.cn/20071228/14/40STATEWORLD00001120.html

下载时请单击右键下载，并将文件后缀名改为.txt

人工智能与数据挖掘实验室

地址：合肥工业大学逸夫楼1208
常小兵首谈联通获WCDMA牌照原因

常小兵表示，中国联通已经具备8年的历史，电信重组后，政府方面也考虑到了联通的现状，所以给了联通WCDMA牌照。

【图】

中国联联通董事长常小兵表示中国联通将充分利用WCDMA牌照的特殊优势，致力于中国联通业务和优势，努力为客户提供优质服务。

【图】

常小兵表示，中国联通已经具备8年的历史，电信重组后，政府方面也考虑到了联通的现状，所以给了联通WCDMA牌照。

【图】

中国联联通董事长常小兵表示中国联通将充分利用WCDMA牌照的特殊优势，致力于中国联通业务和优势，努力为客户提供优质服务。

【图】

常小兵表示，中国联通已经具备8年的历史，电信重组后，政府方面也考虑到了联通的现状，所以给了联通WCDMA牌照。

【图】

中国联联通董事长常小兵表示中国联通将充分利用WCDMA牌照的特殊优势，致力于中国联通业务和优势，努力为客户提供优质服务。

【图】

常小兵表示，中国联通已经具备8年的历史，电信重组后，政府方面也考虑到了联通的现状，所以给了联通WCDMA牌照。

【图】

中国联联通董事长常小兵表示中国联通将充分利用WCDMA牌照的特殊优势，致力于中国联通业务和优势，努力为客户提供优质服务。

【图】

常小兵表示，中国联通已经具备8年的历史，电信重组后，政府方面也考虑到了联通的现状，所以给了联通WCDMA牌照。

【图】

中国联联通董事长常小兵表示中国联通将充分利用WCDMA牌照的特殊优势，致力于中国联通业务和优势，努力为客户提供优质服务。

【图】

常小兵表示，中国联通已经具备8年的历史，电信重组后，政府方面也考虑到了联通的现状，所以给了联通WCDMA牌照。

【图】

中国联联通董事长常小兵表示中国联通将充分利用WCDMA牌照的特殊优势，致力于中国联通业务和优势，努力为客户提供优质服务。
Application 6: Information Fusion with Meta Search

  query = “Xindong Wu Security”
Challenges with Information Fusion

- Intelligent Informatics: Connect seemingly irrelevant information items
  - Whether X is Y’s wife?
  - Did the first ladies meet before?
- Active information fusion
  - Something happens, why?
- Network analysis
  - Sub-network identification
  - Information diffusion
  - Confidence of a node and confidence of fusion
- Adaptive interest modeling and monitoring
  - Sequential pattern mining
  - Event/anomaly detection.
Conclusions

- Ubiquitous personalized information processing involves information aggregation, analysis/mining, user interest modeling, and security/privacy.

- Data mining and sequence matching (with Web information) are 2 research frontiers for ubiquitous personalized information processing.