中国科学院软件研究所学术年会’2017
暨计算机科学国家重点实验室开放周

学术论文

互联网文档流上用户感知的稀有主题序列模式的挖掘方法

Mining User-Aware Rare Sequential Topic Patterns in Document Streams

Jiaqi Zhu, Kaijun Wang, Yunkun Wu, Zhongyi Hu, Hongan Wang

Published in *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 28(7): 1790-1804, 2016.

Contact: zhujq@ios.ac.cn，13683257241

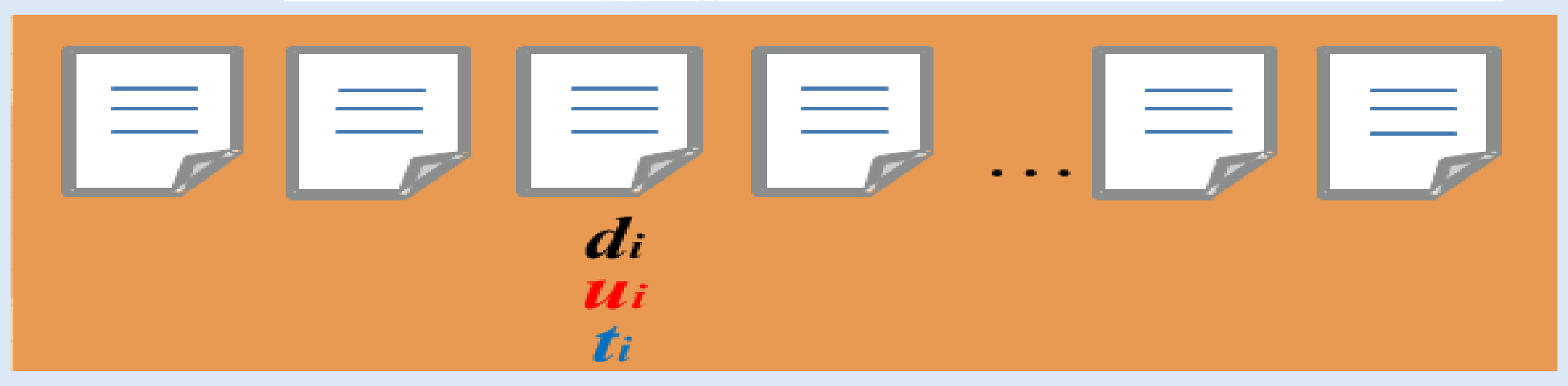
Motivation

Problem: How to discover personalized and abnormal user behaviors from document streams on the Internet?

- to find Internet users with special tasks or goals.
- to detect and infer the real and latent intentions when they publish documents on the internet.

Challenges: These behaviors are complicated with the following characteristics.

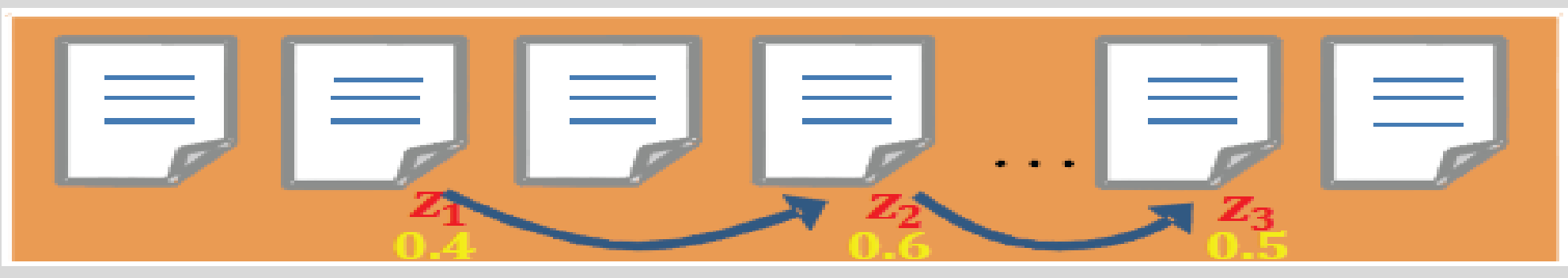
- They should be complete and repeated, so cannot be reflected by one document.
- The texts of documents on the Internet is irregular, so keyword based methods do not work.
- They are generally rare and possibly emergent, so the rules for detecting them are not available.



An example: Lottery fraud via Internet:

- < award temptation → information diddling → fee charging → illegal intimidation >

Contribution



1. We define Sequential Topic Patterns (STP), to build correlation among successive documents of a user.
   - Each of the patterns represents a multiple-step complete behavior, which occurs repeatedly.
   - Topic-level abstract and semantic information is beneficial in finding regularities of users.
   - The probability of topics can be accumulated to achieve high confidence in pattern matching.

Definition (STPs)

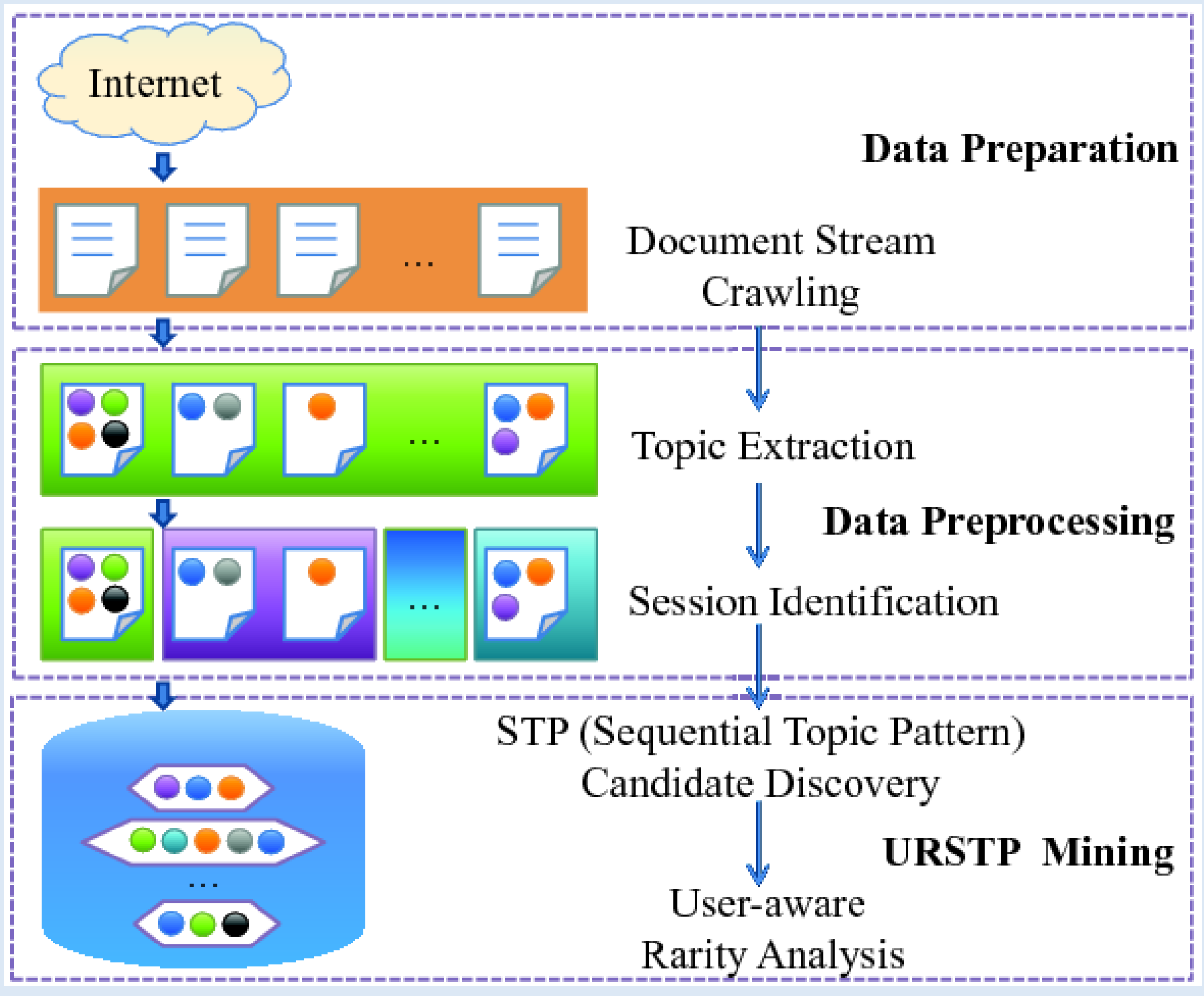
A Sequential Topic Pattern (STP) α is defined as a sequence of topics $< z_1, z_2, \cdots, z_n >$. n is called the length of α.

2. We propose the mining problem of User-aware Rare STPs (URSTPs), to discover personalized and abnormal behaviors of Internet users.
   - User-aware rare: globally rare (for all users), but relatively frequent (for a specific user or a specific group of users).
   - Theoretical significance: define a new kind of patterns for rare event mining, "puts forward a new research direction in Web mining".
   - Practical significance: can be applied in many real-life scenarios of user behavior analysis.
3. We present a framework to solve this problem and design a group of effective and efficient algorithms.

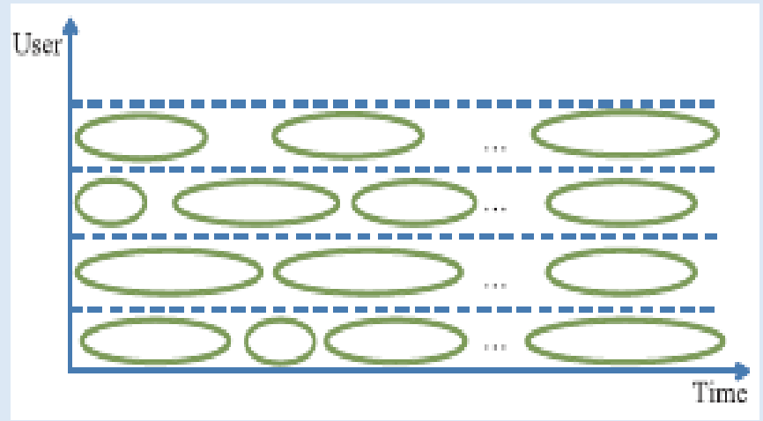
Definition (User-aware Rare STPs)

Given a topic-level document stream TDS, a scaled support threshold h_{ss}, and a relative rarity threshold h_{rr}, an STP α is called a User-aware Rare STP (URSTP) if and only if both $scsupp(\alpha) \leq h_{ss}$ and $RR(\alpha)|_u \geq h_{rr}$ hold for some user u.

Framework and Algorithms



Topic Extraction: TwitterLDA + threshold-based selection
Session Identification: Time interval heuristics



STP Candidate Discovery: to compute scaled supports of all STPs with pattern growth

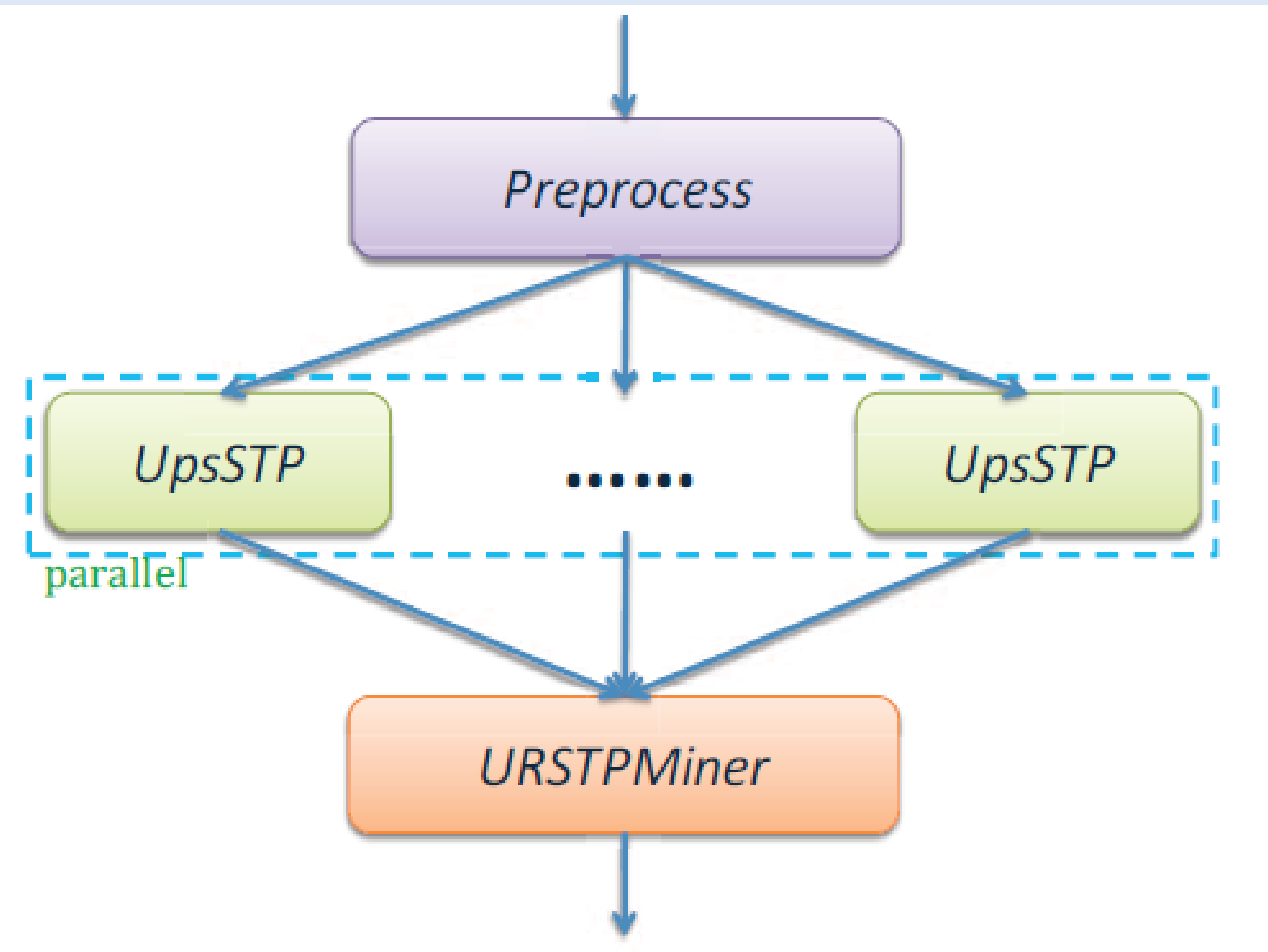
$$supp(\alpha, S) \triangleq \frac{\sum_{i=1}^{|S|} Pr(\alpha \sqsubseteq s_i)}{|S|}$$

$$scsupp(\alpha, S) \triangleq supp(\alpha, S)^{\frac{1}{n}}.$$

User-aware Rarity Analysis:

$$AR(\alpha)|_u \triangleq scsupp(\alpha)|_u - scsupp(\alpha)|_*$$

$$RR(\alpha)|_u \triangleq AR(\alpha)|_u - \frac{\sum_{\beta \in \Phi_u} AR(\beta)|_u}{|\Phi_u|}$$



Experimental Results

Discovered Internet users (top K) are indeed distinctive in real life.

- Compared to approximate ground truth ("Verified" users in Twitter).

Precision	@5	@10	@15	@20	@30
URSTP	0.80	0.70	0.73	0.65	0.60
URSTP-L	0.80	0.60	0.67	0.60	0.53
Topic	0.60	0.50	0.53	0.55	0.47
Topic-L	0.20	0.30	0.33	0.25	0.27

- The users mined by our approach are inclined to be verified users.
- The special behaviors of ordinary users are probably abnormal and should be considered for further investigation.

Discovered STPs are self-interpretable and consistent with tweet contents.

- General Twitter dataset

URSTP	User ID	Scaled support	Relative rarity
(8, 14)	125	0.02	0.318
(13, 2, 8)	207	0.03	0.340

Topic ID	Top words	Description
2	world hours production women goods things oil skin support photo	products
8	buy win fan expensive account mobile care concert ball purpose	buying
13	love health body cool news life pretty skin enjoy great	health
14	day game weekend play happy team class win things amazing	playing

  - User 125 is a sports fan: < buying → playing >
  - User 207 is a cosmetic salesman: < health → products → buying >
- Specific-field Twitter dataset
  - News reporter: < broadcast → NBA players >
  - Ordinary fan: < NBA players → broadcast >

Applications

Economic case investigation: discover new criminals and identify their hierarchy and roles from their communication records.
Financial internal audit: discover and monitor illegal and abnormal operational behaviors of internal staffs in bank.
Credit risk assessment: discover borrowers with latent high risks from their online and offline behaviors.