

Discovering Newsworthy Themes from Sequenced Data: A Step towards Computational Journalism

Abstract:

Today's journalists must pore over large amounts of data to discover attention-seizing facts as news themes. While such a task has traditionally been done manually, there is an increasing reliance on computational technology to reduce human labor and intervention to a minimum. Recently, Zhang et. al. have proposed a new type of news theme, named prominent streaks [4]. A streak corresponds to a set of consecutive sequenced events belonging to the same subject, such as the latest ten games a player participated. The prominent streaks are defined as the skylines among all streaks of a subject, and can be used to represent the history of the subject. In [4], each prominent streak is viewed as a striking news theme since it is outstanding (i.e., nondominated) in a subject's history.

Existing System:

Automatic discovery of newsworthy themes from sequenced data can relieve journalists from manually poring over a large amount of data in order to find interesting news. We study the k-Sketch query processing under both offline and online scenarios. In the offline scenario, our objective is to efficiently discover the sketches for each subject from historical data. A challenging prerequisite is to generate the ranked-streaks. The brute-force approach is to enumerate all streaks to generate ranks. Such an approach has a quadratic complexity wrt. the number of events, which is not scalable even for moderate data size. By leveraging the subadditivity among the upper bounds of streaks, we design two effective pruning techniques to facilitate efficient ranked-streak generation.

Proposed System:

We group consecutive events into streaks and propose a novel idea of ranked-streak to represent the strikingness. We then formulate the k-Sketch query which aims to best summarize

a subject's history using k ranked-streaks. We study the k-Sketch query processing in both offline and online scenarios, and propose efficient solutions to cope each scenario. In particular, we design novel streak-level pruning techniques and a $(1 \pm \epsilon)$ -approximate algorithm to achieve efficient processing in offline. Then we design a $(1 \pm \delta)$ -approximate algorithm for the online sketch maintenance. Our comprehensive experiments demonstrate the efficiency of our solutions and a human study confirms the effectiveness of the k-Sketch query.

Modules:

- Offline k-sketch query processing.
- Online k-sketch maintenance.

SYSTEM REQUIREMENTS

H/W System Configuration:-

Processor	-	Pentium –III
RAM	-	256 MB (min)
Hard Disk	-	20 GB
Key Board	-	Standard Windows Keyboard
Mouse	-	Two or Three Button Mouse
Monitor	-	SVGA

S/W System Configuration:-

Operating System	:	Windows95/98/2000/XP
Application Server	:	Tomcat5.0/6.X
Front End	:	HTML, Jsp

Scripts : JavaScript.
Server side Script : Java Server Pages.
Database : MySQL 5.0
Database Connectivity : JDBC

www.takeoffprojects.com