Research Topics for Graduate Students

Under Continuous Update

Fazli Can
Computer Engineering Department
Bilkent University

March 12, 2009

1. Inverted file creation and update: Suggest a method and implement it and measure its efficiency (Lester, et al., 2006, Lester, et al., 2008)
2. Orhan Pamuk: a stylometric analysis of the Orhan Pamuk novels using machine learning methods (SVM, NN-search, etc.). Possibilities include analyzing the voices of Celal and Galip in *Kara Kitap*, analyzing the narration styles of different novel characters in *Sessiz Ev*, and comparing the narration styles of Faruk Darvınoğlu in *Sessiz Ev* and *Beyaz Kale*.
3. News portal main page news selection: From a set of current news select the ones that would be interesting for news consumers. Compare the performance with human subjects.
4. News portal main page event selection: Assume that we have set of clusters (events) each corresponding a recent topic (event) to be displayed by a news portal. In each cluster we have related news. However, cluster contents are dirty (contain noise/irrelevant data: remember that event detection and tracking is done automatically and non-ideal.). Problems to be solved: which clusters should be invalidated (eliminated), in which order the selected topics should be displayed. (related paper).
5. Opinion retrieval: http://uwspace.uwaterloo.ca/bitstream/10012/4081/1/Thesis_Kun_Cen.pdf
6. Similarity (S) matrix construction: Obtaining the S matrix in an efficient way. Depending on the problem area some non-zero elements of S could be ignored and this would increase the efficiency.
7. Story link detection: determine whether or not two news stories discuss the same topic. One of the tasks of Topic Detection and Tracking research program. (http://projects.ldc.upenn.edu/TDT5/) As a starter see Feng et al. 2008.
8. Stylistic features for Turkish IR: using stylometry for IR (visit http://eprints.sics.se/view/ and look for the publications of Jussi Karlgren, e.g., his dissertations).
9. Translation accuracy measurement: Compare source (original work) and target (translated work) using clustering techniques.
10. Üç İstanbul: hypertextual version of Mithat Kemal Kuntay's novel, some data mining analysis is possible. For some excerpts form the novel see http://www.derkenar.com/kitapkurdu/kuntay.asp, for its TV movie series version information see http://www.sinematurk.com/film.php?7342. For motivation see the hypertext version of Faulkner's *The Sound and The Fury* (http://www.usask.ca/english/faulkner). Another possibility is analyzing the correspondence between the novel and its movie version (see Reyhan Tutumlu's master thesis for some initial ideas, available at http://library.bilkent.edu.tr/, use catalog search to obtain its pdf copy).

Please look at the previous student presentations for understanding the nature of the projects and further possibilities.

References

Ao Feng, James Allan: Finding and linking incidents in news. CIKM 2007: 821-830

Nicholas Lester, Justin Zobel, Hugh E. Williams: Efficient online index maintenance for contiguous inverted lists. Inf. Process. Manage. 42(4): 916-933 (2006). (keywords: in-place, re-merge, reconstruction)

Nicholas Lester, Alistair Moffat, Justin Zobel: Efficient online index construction for text databases. ACM Trans. Database Syst. 33(3): (2008). (keywords: geometric partitioning of inverted indexes)