SPIN: Searching Personal Information Networks*

Soumen Chakrabarti
IIT Bombay

Jeetendra Mirchandani
IIT Bombay

Arnab Nandi
IIT Bombay

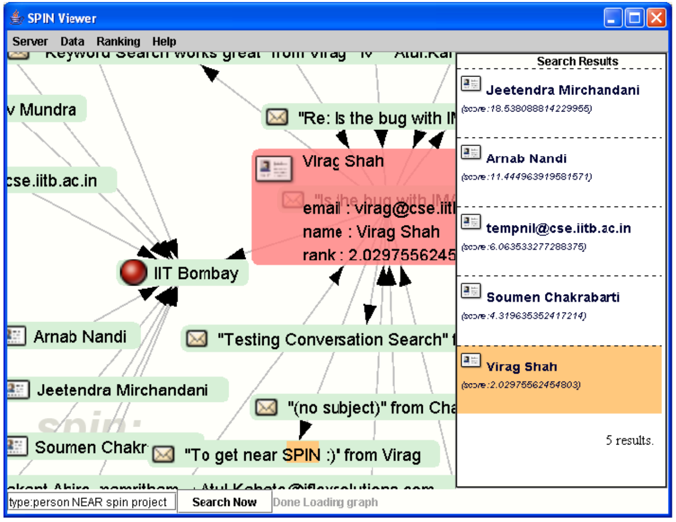
Categories and Subject Descriptors: H.3.3 Information Search and Retrieval: Retrieval models; H.3.4 Systems and Software: Information networks
General Terms: Algorithms, Design, Experimentation
Keywords: Personal information management, information extraction and integration, graphical models for search

1. MOTIVATION

Affordable storage has grown dramatically over the last decade, enabling large-scale archival of email, contacts, documents, images, and music. Large personal storage becomes all the more useful with effective search tools.

Recent operating systems can index PC hard disks and enable keyword search over many file types. All Web-based email services include keyword search. Major Web search companies have recently released desktop search tools that enable search on file systems and browser caches. Many research prototypes exist as well [4, 5, 3, 6].

With some exceptions [3, 6], prototypes and products have stayed close to traditional IR: they largely lack the capability to discover and reconcile *entities* and *relations* that pervade the user's life from diverse and heterogeneous sources, to represent these in a graphical model, and to enable powerful but user-friendly queries on such graphs. These are the goals of our proposed system, SPIN.



email : virag@cse.iit
name : Virag Shah
rank : 2.0297556245

Search Results

Jeetendra Mirchandani
(score:18.538088814229956)

Arnab Nandi
(score:11.444963919581571)

tempnil@cse.iitb.ac.in
(score:6.063533277288376)

Soumen Chakrabarti
(score:4.319635352417214)

Virag Shah
(score:2.02975562454803)

5 results.

type:person NEAR spin project
Search Now
Done Loading graph

2. PERSONAL INFORMATION NETWORKS

Personal Information Network (PIN) entities can be persons, organizations, places, events, projects, trips, software, subscriptions and other artifacts. These are extracted from *mentions* in textual and semistructured sources, such as address books, documents and email.

PIN edges represent relations. Some are "hard" edges explicitly found in the data, e.g., person <u>wrote</u> email or email <u>is-reply-to</u> email. Others are "soft" or probabilistic edges induced through information extraction, e.g., person <u>wrote</u> paper or email <u>mentions</u> person. Yet other soft edges are created by reconciling aliases [7].

3. CAPABILITIES OF SPIN

SPIN provides a proximity-based, type-sensitive and yet schema-light query language to search the PIN. One dominant query paradigm is `type=`*TypeName* `NEAR` *Predicates*. The user can look for a student who graduated around 2001 and went to work at IBM using the query `type=person NEAR org=IBM year=2001`. The predicates `org=IBM` and `year=2001` *activate* some nodes, which spread the activation using algorithms [1] that are sensitive to the structure and uncertainty in the PIN. SPIN can also search for small PIN subgraphs that connect at least one node matching each query keyword [2]. Moreover, SPIN will allow *situated* visual queries where the user can drag and drop PIN nodes, or designate nodes as *hot*. SPIN continuously assimilates user edits and annotations into its ranking algorithms. It also has user-trainable Web explorers and extractors to *augment* the PIN semi-automatically. Finally, SPIN serves as a type-aware middleware that uses the PIN and its trained type recognizers to enhance queries to, and filter responses from, typeless keyword search engines on the Web.

4. REFERENCES

[1] A. Balmin, V. Hristidis, and Y. Papakonstantinou. Authority-based keyword queries in databases using ObjectRank. In *VLDB*, Toronto, 2004.

[2] G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, and S. Sudarshan. Keyword searching and browsing in databases using BANKS. In *ICDE*. IEEE, 2002.

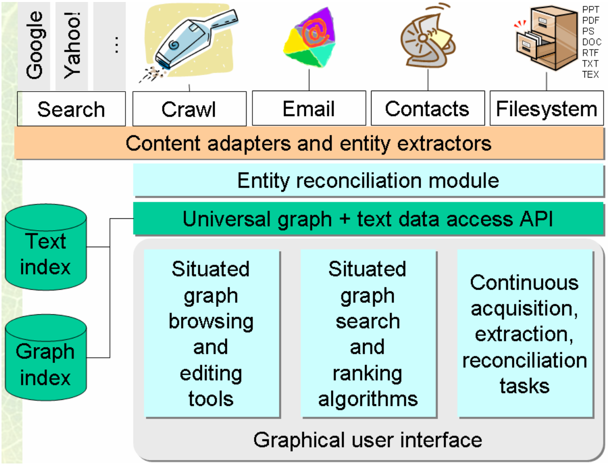
[3] X. Dong and A. Y. Halevy. A platform for personal information management and integration. In *CIDR*, 2005.

[4] S. T. Dumais, E. Cutrell, J. J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. Stuff I've Seen: A system for personal information retrieval and re-use. In *SIGIR*, 2003.

[5] J. Gemmell, G. Bell, R. Lueder, S. Drucker, and C. Wong. MyLifeBits: Fulfilling the Memex vision. In *ACM Multimedia*, pages 235–238, 2002.

[6] D. Quan, D. Huynh, and D. R. Karger. Haystack: A platform for authoring end user semantic web applications. In *International Semantic Web Conference*, 2003.

[7] B. Wellner, A. McCallum, F. Peng, and M. Hay. An integrated, conditional model of information extraction and coreference with application to citation matching. In *UAI*, 2004.


Google
Yahoo!
:
PPT
PDF
PS
DOC
RTF
TXT
TEX
Search
Crawl
Email
Contacts
Filesystem
Content adapters and entity extractors
Entity reconciliation module
Universal graph + text data access API
Text index
Graph index
Situated graph browsing and editing tools
Situated graph search and ranking algorithms
Continuous acquisition, extraction, reconciliation tasks
Graphical user interface

*See `http://www.cse.iitb.ac.in/~soumen/` for details.

SIGIR'05, August 15–19, 2005, Salvador, Brazil.
ACM 1-59593-034-5/05/0008.