



In the NEWS

ALSO FEATURED THIS ISSUE

**PHONETIC-SEQUENCE ALGORITHM
COULD REDUCE DRUG NAME CONFUSION**

Features Editor: **Dennis Taylor**
dtaylor@computer.org

AI and Privacy on the Web

Danna Voth

The Web's inherent openness and potential for growth and new applications make it a powerful tool, especially as methods develop for handling data on the Semantic Web. (In the Semantic Web, heterogeneous data sources can be treated as a single source, with an ontology that mediates between them, thereby allowing the creation of knowledge out of information.) But formidable problems accompany its potential to publish and store

massive amounts of information. As more users access and reason over the information on the Web, particularly in pursuit of creating more valuable knowledge, issues arise of securing and protecting personal information and of this information's accuracy and provenance. Computer scientists are addressing these problems, imagining and developing solutions that range from semantic firewalls to methods for monitoring and auditing the use—instead of flow—of information.

Privacy policy problems

A major problem with some AI approaches to Web privacy, says Internet security consultant Rebecca Herold, is that their use might be compromised owing to a lack of common terminologies, particularly on the Semantic Web. "The fact that there's not a common language built around how the same type of information from one entity to another is being handled could lead to some very huge mistakes in interpreting that data," she says.

Herold points to the recent US National Security Agency handling of telephone

records. "You have all these phone records, and the different data fields within those phone records might be labeled in many different ways from one phone company to the next," Herold says. "But when you start trying to analyze them and interpret them, it would be like having somebody speaking French and somebody speaking Italian. You're going to come up with some miscommunications that can lead to some information which won't be correct."

Even when suspicious Internet activity can be identified, an instance of it might be misinterpreted. Many sites, such as those providing pornography, make themselves easy to find by incorporating search terms similar to those of other common destinations. Internet users can easily stumble across them without meaning to. "If you had a sophisticated AI system that knew it [a pornography site] was a bogus site," Herold says, "it might incorrectly interpret the activities of people who don't know what these sites are about."

Coping with more

Hugh Glaser, a University of Southamp-

ton reader in computer science, views the Web's potential as a challenge to both use and abuse. "From the AI point of view, with every challenge it's an arms race," he says. The maintenance of anonymity on the Web is such a challenge. "The traditional ways of anonymizing personal data assume the data is not being correlated with all sorts of other resources," he says. "In the Semantic Web, where you can more effectively correlate between lots of resources, those techniques break down."

Data mashing, or combining multiple database sources to create specific, individualized information, can be a potential means of de-anonymizing on the Web. Glaser imagines a scenario where different data comes from two separate sources, both of whom are careful to anonymize the data. For example, census data, including age and location, is released on the Web. Epidemiological data, correlated by age, is also published. "It is the combination that facilitates the de-anonymization," Glaser says. Depending on various statistical parameters, you could then identify individuals and their illnesses. "The extra dimension of the Semantic Web is that the combination is more easily achieved when dealing with knowledge than raw data."

Such revelation of personal data through de-anonymizing can present privacy issues. The Semantic Web's potential to dynamically create knowledge out of various sources of data could inspire what Glaser calls a "semantic firewall," a policy that allows inference of knowledge without revealing personal data. "At that stage it becomes a more intelligence-based activity," Glaser says. For example, when rent-

ing a car, instead of providing your birth date, you would give the rental company just enough information for it to determine you're over 21.

Trust solutions

The amount of information collected, distributed, and replicated on the Web is difficult to control, notes Lalana Kagal, postdoctoral associate at MIT's Computer Science and Artificial Intelligence Laboratory. As people mine information posted online from social networks, public records, commercial databases, and blogs, Kagal says that they could "infer conclusions that could be incorrect and potentially harmful." She proposes that computational policies involve trust agreements to protect privacy concerns. Information users would agree to access and use the information under certain conditions, and suffer institutional and legal consequences for misuse. "Nowadays most of privacy research is focused on the server side and on anonymization techniques," Kagal says. "We want the users to have more control."

Rei, a policy language that reasons over Semantic Web descriptions (<http://rei.umbc.edu>), can be placed on individual devices or embedded into the infrastructure to monitor how information is shared or mined. Kagal says that the Policy Aware Web project (<http://policyawareweb.org>), a developing framework for dynamic Web use, promotes the development of such technologies that help users "define trust-based policies in their policy languages and over their own domain information." An example of such a policy used in a medical environment would allow specialists to access information on patients only if they have previously consulted on the specific cases.

Transparency approach

Daniel J. Weitzner, director of the World Wide Web Consortium's Technology and Society Activities and a research scientist at MIT's CSAIL, wants to move the focus on Web privacy issues from controlling data to monitoring and accounting for its uses. "In this environment, where there are going to be very complex inferences done which involve people's personal information, we want to give people a mechanism whereby they can look at inferences, identify the ground facts of personal information on which those inferences are based, and be able to tell whether that information about them is accurate," he says. The transparency

of the reasoning operations over the multiple sources of personal information available from multiple databases can provide accountability of the information's use. However, legal and social rules must first be established as the basis for this accountability and for the consequences of illegal or inappropriate uses.

Building on top of the developing Semantic Web infrastructure, Weitzner and his team are using AI technologies to create an inferencing engine, a truth maintenance system, and a proof generator. They're trying to understand what types of laws would be required to protect privacy, as well as the kind of expressivity and form of rules that would be required. "What we are imagining in a data mining context is that there would be some investigation which would be conducted by criminal-law-enforcement agents who would be working with data generating inferences about people. An inference might include a conclusion that a certain person is suspicious and should be stopped at an airport for further screening," Weitzner says. "What we add to that is a truth maintenance system that will keep track of the chain of inferences, keep track of all the information that has been used to generate that final conclusion of stopping that person at the airport."

The truth maintenance system will be able to provide both the transparency into the information used and that user's accountability to a set of rules. When an action occurs owing to a conclusion inferred from the data, the proof generator can try to show that the inference was premised on a reasonable set of assumptions that comply with the rules governing that information's use. The proof generator would have to collect the normal evidence gathered when someone is indicted for a crime, such as details indicating that that person is likely guilty. It would also generate a proof that the information is used correctly. "Putting these all together, we have an environment that not only supports normal law enforcement and national-security investigative activity," Weitzner says, "but does so in way that it carries with it a kind of auditable trail that allows one to establish either compliance or lack of compliance with the set of rules."

Expanding scope

Weitzner describes the challenge of coming up with new approaches to privacy on the Web: "What makes you think you can



IEEE Computer Society Publications Office

10662 Los Vaqueros Circle, PO Box 3014
Los Alamitos, CA 90720-1314

STAFF

Lead Editor
Dennis Taylor
dtaylor@computer.org

Group Managing Editor
Crystal R. Shif
cshif@computer.org

Senior Editors
Shani Murray, Dale Strok, and Linda World

Assistant Editor
Brooke Miner

Editorial Assistant
Molly Mraz

Magazine Assistant
Hilda Carman

Contributing Editors
Annette Ibrahim and Joan Taylor

Design Director
Toni Van Buskirk

Layout/Technical Illustrations
Carmen Flores-Garvey and Alex Torres

Publisher
Angela Burgess, aburgess@computer.org

Associate Publisher
Dick Price

Membership/Circulation Marketing Manager
Georgann Carter

Business Development Manager
Sandra Brown

Senior Production Coordinator
Marian Anderson

Submissions: For detailed instructions and formatting, see the author guidelines at www.computer.org/intelligent/author.htm or log onto *IEEE Intelligent Systems'* author center at Manuscript Central (www.computer.org/mc/intelligent/author.htm). Visit www.computer.org/intelligent for editorial guidelines.

Editorial: Unless otherwise stated, bylined articles as well as products and services reflect the author's or firm's opinion; inclusion does not necessarily constitute endorsement by the IEEE Computer Society or the IEEE.

How to Reach Us

Writers

For detailed information on submitting articles, write for our Editorial Guidelines (isystems@computer.org) or access www.computer.org/intelligent/author.htm.

Letters to the Editor

Send letters to

Dennis Taylor, Lead Editor
IEEE Intelligent Systems
10662 Los Vaqueros Circle
Los Alamitos, CA 90720
dtaylor@computer.org

Please provide an email address or daytime phone number with your letter.

On the Web

Access www.computer.org/intelligent for information about IEEE Intelligent Systems.

Subscription Change of Address

Send change-of-address requests for magazine subscriptions to address.change@ieee.org. Be sure to specify IEEE Intelligent Systems.

Membership Change of Address

Send change-of-address requests for the membership directory to directory.updates@computer.org.

Missing or Damaged Copies

If you are missing an issue or you received a damaged copy, contact membership@computer.org.

Reprints of Articles

For price information or to order reprints, email isystems@computer.org or fax +1 714 821 4010.

Reprint Permission

To obtain permission to reprint an article, contact William Hagen, IEEE Copyrights and Trademarks Manager, at copyrights@ieee.org.

do any kind of useful reasoning in this open world environment? It's fine if you've got a limited number of databases or a limited number of information sources and a fixed set of rules. You can certainly imagine that it would be possible to reason in a finite way over that data and those rules." But, he says, "we don't know at the beginning the

exhaustive description of all the data we're dealing with. We don't even necessarily know at the beginning of an inferencing process an exhaustive description of all the rules we're dealing with, since laws change and circumstances change. Our hardest problem is that we're really trying to provide accountability to rules at Web scale."

Phonetic-Sequence Algorithm Could Reduce Drug Name Confusion

Jan Krikke

Every year thousands of people in the US die or are seriously harmed because of medical errors. Many of the fatalities are due to medication name mix-ups (for example, Amaryl/Amikin, Flomax/Volmax, and Verelan/Virilon). The US Food and Drug Administration, responsible for approving new drug names, tries to mitigate the problem with Phonetic Orthographic Computer Analysis, a system for computationally comparing look-alike or soundalike drug names. At the heart of POCA is ALINE, an algorithm for aligning phonetic sequences developed by Greg Kondrak of the University of Alberta. ALINE is unique because it can deal with how words are pronounced in any language.

A 1999 Institute of Medicine report (*To Err is Human: Building a Safer Health System*) first highlighted the problems with confusing drug names. The report claimed that preventable medical errors in US hospitals result in the death of 44,000 to 98,000 people every year. An estimated 12.5 percent of these errors result from confusion about drug names. In 2002, the FDA decided to automate evaluation of new drug names, using phonetic-alignment algorithms and other natural language processing technologies. The agency experimented with several programs with phonetic search capabilities but didn't achieve acceptable results until they tested ALINE.

How ALINE works

ALINE (a demo is available at www.cs.ualberta.ca/~kondrak/cgi-bin/demo/aline/aline.html) combines existing sequence-comparison techniques with a scoring scheme to compute phonetic similarity. (The program requires words to be represented by phonetic symbols.) Written in C++, it takes into account parts of the vocal tract (nasal cavity, pharynx, vocal cords, and so on), types of articulation (bilabial, palatal, uvular, and so on), and other dimensions of human speech such as aspiration. In "Automatic Identification of Confusable Drug Names" (Jan. 2006 *Artificial Intelligence in Medicine*), Kondrak and coauthor Bonnie Dorr explain that ALINE's principal component is a function that calculates two phonemes' similarity.

"Phonemes are expressed in terms of binary or multivalued phonetic features," Kondrak and Dorr write. "For example, the phoneme *n*, which is usually described as a *voiced alveolar nasal stop*, has the following feature values: *Place* = 0.85, *Manner* = 0.6, *Voice* = 1, and *Nasal* = 1, with the remaining features set to 0. In order to compute the phonetic distance between two phonemes, the differences between their numerical values for each feature are multiplied by the feature's salience weight ..., and the resulting values are summed up." They go on to say that ALINE then calculates the phonetic-similarity score by subtracting the distance from the maximum score. To emphasize consonant correspondences, ALINE decreases the similarity score further if one or both of the phonemes are vowels.

Kondrak and Dorr demonstrated that

ALINE outperforms orthographic approaches on a test set containing more than 2,000 soundalike confusion pairs. But they also concluded that a combination of phonetic and orthographic measures produces the best results.

From linguistics to healthcare

Kondrak, who is fluent in three languages and proficient in another three, initially developed ALINE to help linguists find similarities between words when searching for language histories. The program is part of the Upper Necaxa Totonac Project (www.arts.ualberta.ca/~totonaco). "Upper Necaxa is a seriously endangered language, which is still spoken by a few thousand indigenous people in the Puebla State in Mexico," Kondrak explains. "The primary goal of the project is to document the language through the compilation of an extensive dictionary and other resources, which may aid the revitalization efforts. My software is used for the identification of cognates and regular sound correspondences within that language family."

In 2003, ALINE caught the attention of researchers at the Project Performance Cor-

poration, which the FDA had contracted to develop POCA. Although ALINE was developed for language histories, Kondrak says that you can easily adapt it to other requirements. He explains: "ALINE has a number of parameters that can be adjusted, or tuned. We tuned the parameters using a large set of drug name pairs that had been reported as confusable."

A complete solution remains elusive

ALINE is likely to help prevent future approval of confusing pairs like Zantac/Xanax and Verelan/Virilon, but the FDA's Laura Alvey points out that POCA is only one tool the FDA uses to assess trade names. (For example, the FDA also uses various databases, including a United States Pharmacopeia database that records deaths and injuries caused by drug name confusion.) "It is not used as the decision maker for approval," she explains. "POCA offers a quicker way to search for similar sounding names. POCA's utility is directly linked to the data sources it can search against, which is limited at this time." Moreover, POCA won't deal with existing problems.

An estimated 600 drugs with similar names are on the market. (One of the most frequently confused pairs is Primaxin, an antibiotic injection, and Primacor, a hypertension injection. Confusing them can result in death.)

The question remains why the medical establishment is only now addressing this issue with computerized systems. The required technology has been available for at least two decades. Not surprisingly, several US hospitals have taken matters into their own hands. At Memorial Sloan-Kettering Cancer Center in New York City, all doctors must prescribe drugs via computer. All medications are digitally double-checked. The US Veterans Administration has installed computer bar code medication administration technology at all its medical facilities. The system aims to ensure that the right medication is given to the right patient at the right time. Medical staff uses handheld scanners to compare bar codes on the medication with bar codes on the patient's wristband. Mandating such technologies would not only save lives, it would reduce the enormous financial cost of dealing with medical errors. ■

Special Issue on Intelligent Educational Systems

C
A
L
L

F
O
R

P
A
P
E
R
S

Submission deadline: 17 Nov. 2006



Publication: July/August 2007

Education in the 21st century has to handle distributed content and geographical dispersion of students and teachers. This raises new issues regarding forms of instructional interaction, affecting nature of learning processes, such as distance learning, lifelong education, and on-the-job training. This poses strong demands for intelligent educational systems tailored to both students' and teachers' individual needs. For this special issue, we invite papers that discuss novel methods, tools, and applications addressing this field's key challenges such as

- ✧ semantic interoperability, both of subject-domain data and instructional data,
- ✧ context-sensitive feedback generation
- ✧ alignment of teacher and student viewpoints
- ✧ personalized content delivery and generation, and
- ✧ motivational and affective interactions.

We expect papers to report on case studies, empirical research, and/or real-world systems that illustrate novel applications of such technologies that can bring the state-of-the-art in intelligent education systems to a new usage level.

Submission Guidelines: Manuscripts should be 3,000 to 7,500 words (counting a standard figure or table as 200 words) and should follow the magazine's style and presentation guidelines (see www.computer.org/intelligent/author.htm). References should be limited to 10 citations. To submit a manuscript, access the IEEE Computer Society Web-based system, Manuscript Central, at <http://cs-ieee.manuscriptcentral.com/index.html>. For the full call, see www.computer.org/portal/pages/intelligent/content/educfp.html.



Guest Editors:

Lora Aroyo
l.m.aroyo@tue.nl

Art Graesser
a-graesser@memphis.edu

Lewis Johnson
johnson@isi.edu