

Document and character recognition in the Open Mind Initiative



David G. Stork

Ricoh Silicon Valley

stork@OpenMind.org

ICDAR99, Bangalore India

Outline

- One-sentence description
- Background
- Open Mind Initiative
- Document & character projects
- Relation to Open Source and to Data mining
- New problems and opportunities
- The future...

Open Mind Initiative

A collaborative framework (based on Open Source methodology) for developing “intelligent” software, where...

- » domain experts provide algorithms,
- » tool developers provide software infrastructure and tools, and
- » non-expert ‘netizens’ provide raw data over the web.

Background:

E-community & Open Source Waves

- Free Software Foundation
 - » GNU, gcc, emacs, ... are high-quality and very stable
- Linux
 - » 10M lines; 10M seats; dbl. time 6 mo., 10^5 contributors
- Newhoo!
 - » Open web directory (540,000 sites; 11,000 editors; 83,000 categories)
- Infomedia
 - » Open source encyclopedia

Growth of new software methods

1990 10^5 programmers

1995 10^6 web authors

1999 10^9 netizens

1995 Linux

1999 Newhoo!

2003 Open Mind

- New communication allows communities and collaboration, and thus new software methods
- Opportunities expand to less-skilled users

Background: Pattern recognition/ intelligent systems

- Recognizer = **Theory** + **Model** + **Data**
- **Theory**: excellent
 - Adequate for building useful systems
- **Models**: depend on problem
 - need framework for rapid experimentation/
collaboration
- **Data**: can always use more!
 - » “the group with the most data wins”

Open Mind Initiative

- Three main functions provided by
 - » Domain Experts
 - fundamental algorithms, process control, education/proselytizing, ...
 - » Tool developers
 - software infrastructure, tools, ...
 - » Netizens
 - raw data, low-level bug reports, ...

Domain Experts

- Provide algorithms (e.g., OCR, document recognition...)
- Provide general algorithms (e.g., Bayes nets, ...)
- Process control
 - » detect outliers for review/rejection
 - » data “voting”
 - » catch trials
 - » signal dection theory (d')
 - » method of limits
 - » bias avoidance
 - » two-alternative forced-choice hidden staircase
- Trend to publish data and algorithms on the web
- Open framework allows rapid exploration of models

Tool/infrastructure developers

- Get maximum information for minimum netizen effort (e.g., informative patterns)
- Make it easy (fast) for contributors
- Reward contributors
- Web infrastructure
- Collaborative software (version control)

Netizens

- Incentives
 - » benefits in used system
 - » fun (games: Marathon, MUDD, ...)
 - » recognition (post names by amount of info. accepted)
 - » general interest (note progress: data and performance)
 - » altruism/philanthropy (cf. OED, SETI, ...)
 - » education (linguistics in schools, ...)
 - » lottery
 - » frequent flyer miles
 - » money
- Web allows **many** netizens to contribute
 - » 1.5M inmates, 1M in nursing homes, ...

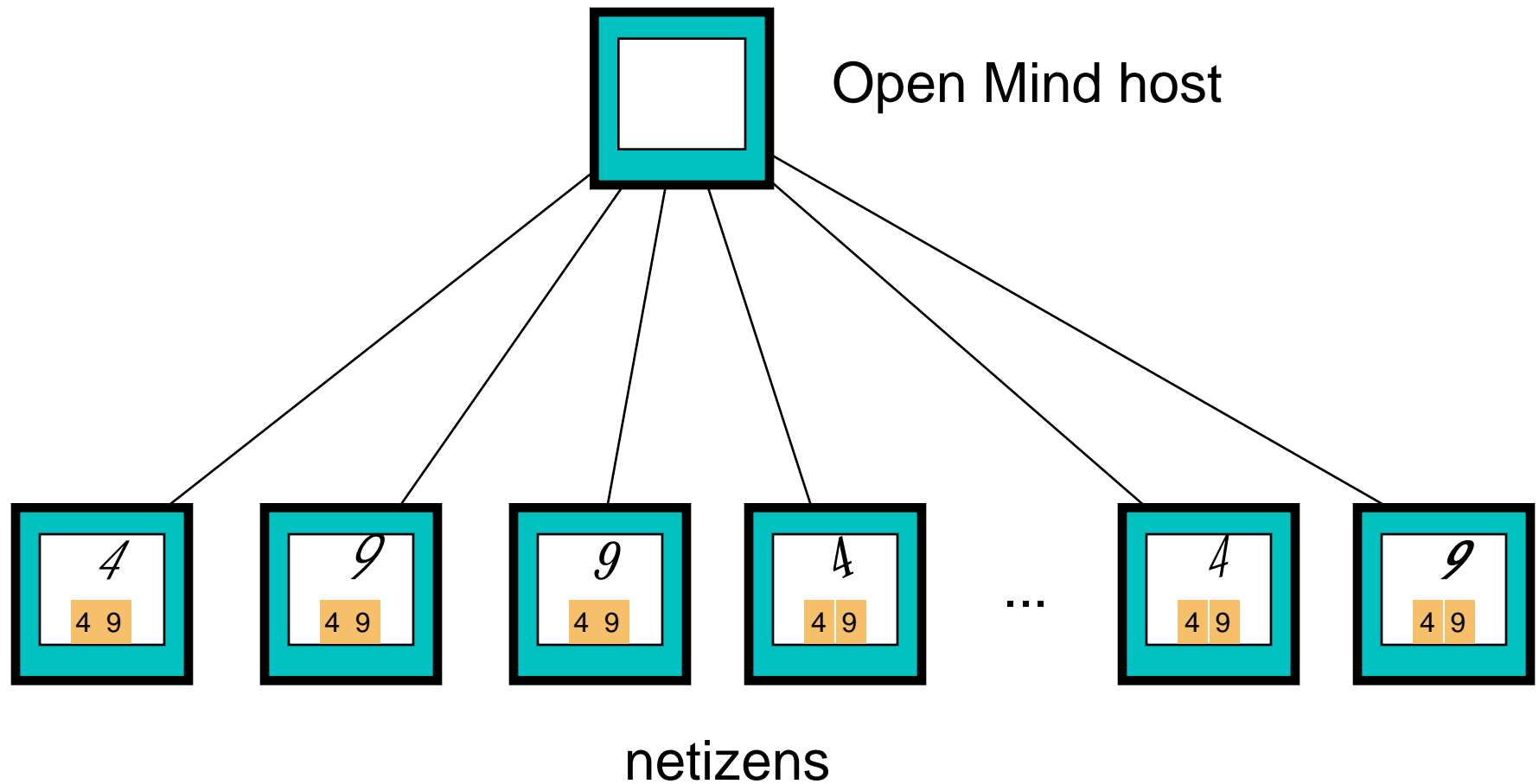
Sample Projects (1)

11

Isolated handwritten character recognition

- Recognizer: simple neural net, decision tree, nearest-neighbor, ...
- Patterns presented on netizens' browsers, cached, ...
- Synthetic data (rotate, skew, line thicken/thin, ...)
- Learning with queries (ask informative patterns); each pattern more valuable than a sampled one
- Cooperative improvement (submit characters over internet, download improved OCR the next day)
- game interface
- Improved OCR

OCR example

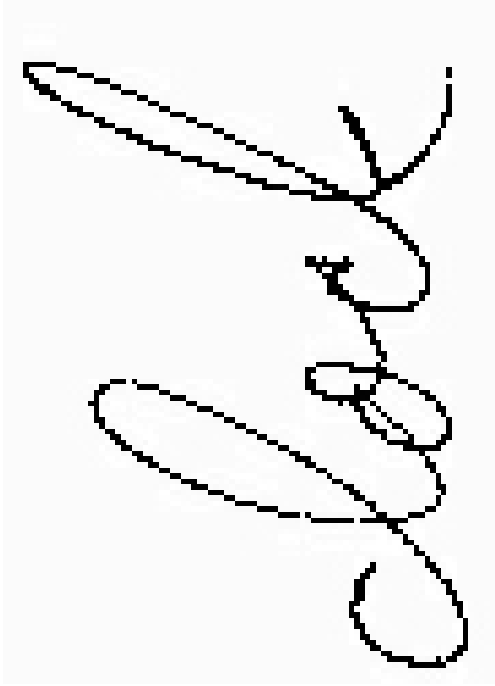


Sample Projects (2)

Handwritten word recognition

- Recognizer: “from the literature”
- Words scanned from handwritten docs, possibly with one or two context words
- Three alternatives shown, best selected by netizen (as in commercial speech recognizers)
- Improved handwritten word recognition

Handwritten word recognition



clock

dock

clark

...

Sample Projects (3)

Open Mind spam filter

- Netizens (motivated!) forward e-mail spam to Open Mind host
- algorithms learn **content**: keywords, phrases, frequencies, ...
- linguistic models refined by labeled spam and non-spam messages
- **Improved spam filter**

Sample Projects (4)

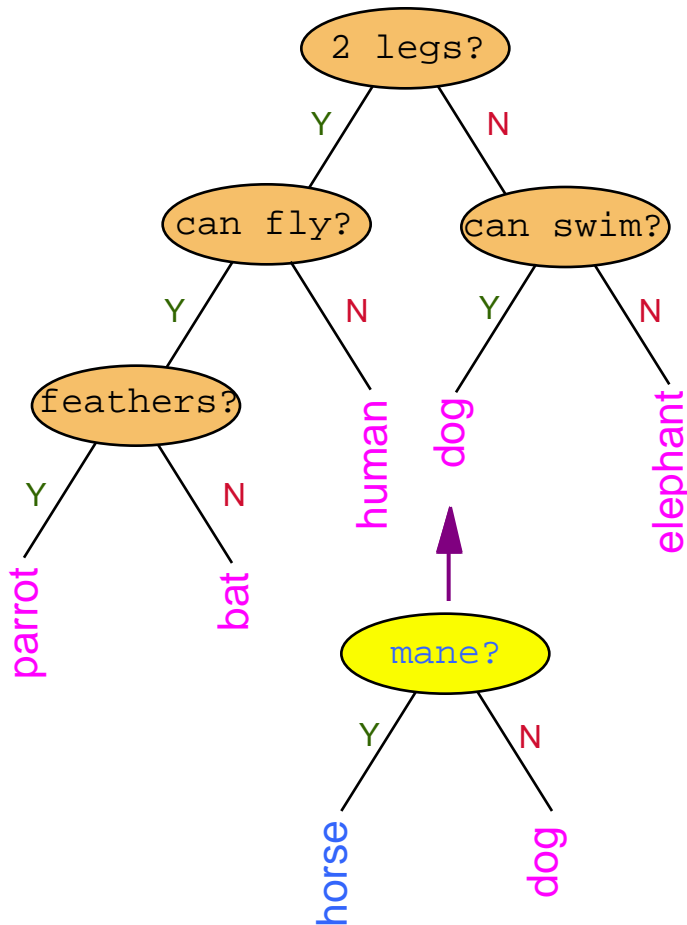
Open Mind chatbot

- MUDD-like game
- Goal: find the route through the castle to the “human”
 - » choose the “most natural” paragraph
- Linguistic information learned in background
- More natural interfaces; Loebner Prize

Sample Projects (5)

17

Open Mind Animals (Lam & Stork 99)



challenges:

truthing

- insure valid animals & data quality
- name/synonym check

bug reporting

- forwarding errors to domain experts

crediting contributors

- ordered by amount contributed
- avoid ID clashes; allow anonymity

query simplification

- reduce average number of queries/new animal

tree simplification

- better taxonomy

arbitrary branching factor

- tree reflects the structure of domain

coverage

- ensure "all" animals appear (enter "ibex")

generalizable to other domains

- other forms of queries

human-machine interface

- natural, show current query set (selectable)

display progress

- number of animals, contributors, tree itself

Open development allows cross-project integration

- Use Open Mind speech as front end for games used in other projects
- Use Open Mind linguistics for Open Mind OCR
- Use Open Mind common sense for Open Mind linguistics

New theoretical problems

- Data truthing and outlier detection
- Relative value of learning with queries vs. iid samples
- Optimal learning strategies given...
 - » Bayes error
 - » probability of hostile data
 - » probability of data error
- Learn reliability of netizens, individually and as a group
- Scalable algorithms

Relation to Open Source

Open Source

- no netizens
- expert knowledge (C++filt,gdbm)
- machine learning irrelevant
- web infrastructure useful
- most work is directly on the final software
- hacker culture (10^5)

Open Mind

- netizens crucial
- informal knowledge (read, hear)
- machine learning essential
- web infrastructure essential
- most work is on the infrastructure
- netizen and business culture (10^9)

Relation to Data Mining

Data Mining

- type of data may not be available for the project desired (e.g., OCR)
- no interactive queries
 - slower learning
 - ambiguities not resolved
- relatively fixed amount of data
- model data **as it exists**
- little or no netizen support

Open Mind

- data tailored to the project desired (e.g., OCR)
- interactive queries
 - faster learning
 - ambiguities resolved
- new data encouraged
- collect data for **best classifier**
- netizen support

Related efforts elsewhere

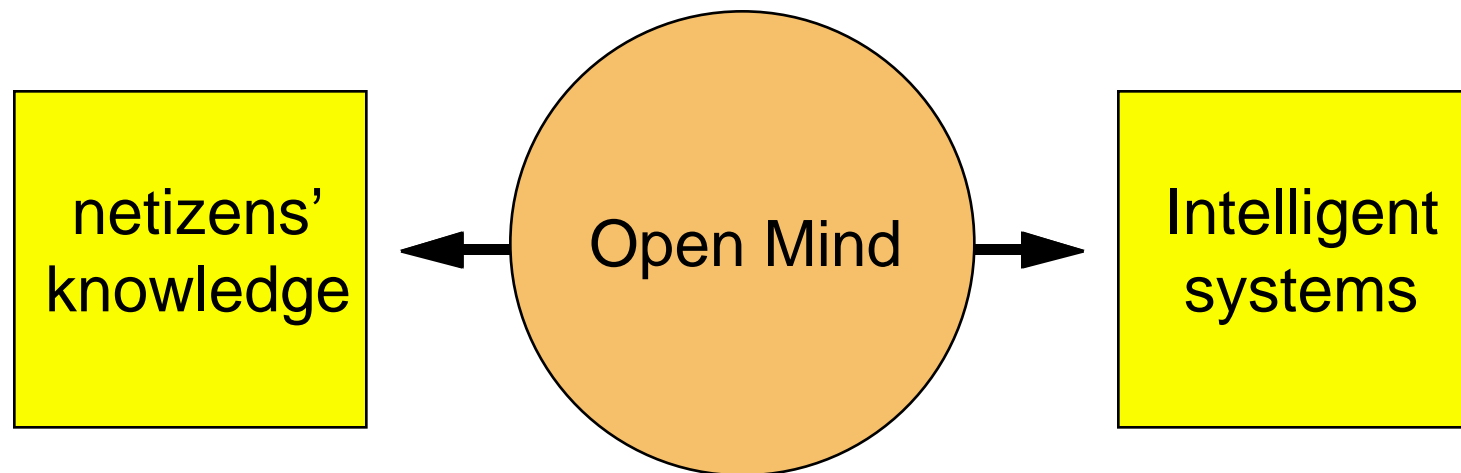
- Speech
 - » Linguistic Data Consortium
 - » Macrophone
 - » Human phoneme project
 - » CSLU (Center for Spoken Language Understanding) Open Source speech tools
- OCR
 - » NIST, CEDAR, ARPA, UNIPEN, ...
- GNU dictionary
- Newhoo!

It is inevitable

Need is here

Web is here

Theory/Machine learning is here



This collaboration is going to happen!

- » Less radical than Richard Stallman or Linus Torvald...

Status

- www.OpenMind.org now up
- Open Mind Animals being bulletproofed
- Logo (Imageworks, Inc.)
- Website design (Diamond Bullet, Inc.)
- Public relations (Ruder Finn, Inc.)
- Corporate donations explored (e.g., books, CDs, ...)
- 20 lectures, 4 publications...

Committed projects



University of Sherbrooke
(Canada)



MIT



Nijmegen University
(Netherlands)

Summary

- Open Mind
 - » Collaborative framework for developing “intelligent systems”
 - » Experts, tool developers, netizens
- Projects
- Vision of the future

Take home messages

- **Open software development**
 - » leads to high-quality software
 - » integrate components
 - » can be used for many projects in document understanding
- **Netizens**
 - » can contribute large amounts of informal data
- **Very large data sets**
 - » needed for further progress
- **Data collection**
 - » opportunities for algorithms and theory

Questions/Comments...



“building intelligent systems one netizen at a time”

www.OpenMind.org