Collective Intelligence and Machine Learning

Haym Hirsh
Professor, Department of Computer Science, Rutgers University
Visiting Scholar, Center for Collective Intelligence, MIT
hirsh@cs.rutgers.edu
Collective Intelligence

Swarm Intelligence
Collective Intelligence
Swarm Intelligence
Hive Intelligence
Collective Intelligence
Swarm Intelligence
Hive Intelligence
Natural Computation
Particle Swarm Optimization

James Kennedy\textsuperscript{1} and Russell Eberhart\textsuperscript{2}

\textsuperscript{1}Washington, DC 20212\newline kennedy\_jim@bls.gov

\textsuperscript{2}Purdue School of Engineering and Technology\newline Indianapolis, IN 46202-5160\newline eberhart@engr.iupui.edu

\section*{ABSTRACT}
A concept for the optimization of nonlinear functions using particle swarm methodology is introduced. The evolution of several paradigms is outlined, and an implementation of one of the paradigms is discussed. Benchmark testing of the paradigm is described, and applications, including nonlinear function optimization and neural network training, are proposed. The relationships between particle swarm optimization and both artificial life and genetic algorithms are described.

\section*{1 INTRODUCTION}
Natural Computing

An International Journal
Editors-in-Chief: G. Rozenberg; H.P. Spaink
ISSN: 1567-7818 (print version)
ISSN: 1572-9796 (electronic version)
Journal no. 11047
Exercise 1
Exercise 2
Hi you all:

I don't know if this is the proper place to post this, but I got the message below from one of the other newsgroup I'm in, and I just want to post it here to see if anyone could help.

---------- forwarded message begins ---------------
Date: Mon, 10 Apr 1995 04:44:56 PST
Reply-To: Medical student discussion list <MEDSTU-L%UNMVMA.BIT_@cmsa.Berkeley.EDU>
Sender: Medical student discussion list <MEDSTU-L%UNMVMA.BIT_@cmsa.Berkeley.EDU>
From: Cai Quanqing <ca_@MCCUX0.MECH.PKU.EDU.CN>
X-To: medst_@unmvma.bitnet
To: Multiple recipients of list MEDSTU-L <MEDSTU-L%UNMVMA.BIT_@cmsa.Berkeley.EDU>

Hi,

This is Peking University in China, a place those dreams of freedom and democracy. However, a young, 21-year old student has become very sick and is dying. The illness is very rare. Though they have tried, doctors at the best hospitals in Beijing cannot cure her; may do not even know what illness it is. So now we are asking the world -- can somebody help us?

Here is a description of the illness:

The young woman -- her name is Zhu Ling -- is a student in the chemistry department. On DEC. 5, 1994, Zhu Ling felt sick to her stomach. Three days later, her hair began to fall out and within two days she was completely bald. She entered the hospital, but doctors could not discover the season for her illness. However, after she was in the hospital for a month, she began to fell better and her hair grew back. Zhu Ling went back to school in February, but in March her legs began to ache severely, and she felt dizzy. She entered XieHe Hospital - Chinese most famous hospital. In early March and on March 15, her symptoms worsened. She Began to facial paralysis, central muscle of eye's paralysis, self-controlled respiration disappeared. So she was put on a respirator.

The doctors did many tests for many diseases(include anti-H2V, spinal cord puncture, NMR, immune system, chemical drug intoxication ANA,ENA,DSONA,ZG and Lyme), but all were negative, except for Lyme disease(ZGM(+)).
The doctors now think that it might be acute disseminated encephalomyelitis (ADEM) or lupus erythematosus (LE), but the data from the tests do not support this conclusion.

The doctors are now treating Zhu Ling with broad-spectrum antibiotic of cephalosporin, anti-virus drug, hormone, immunoadjuvent, gamma globulin intravenous injection and have given her plasma exchange (PE) of 10,000 CCs. But Zhu Ling has not responded -- she reamers in a vegetative state, sustained by life support.

If anyone has heard of patients with similar symptoms -- or have any ideas as to what this illness could be, please contact us. We are Zhu Ling's friends and we are disparate to help her.

This is the first time that Chinese try to find help from Internet, please send back E-mail to us. We will send more crystal description of her illness to you.

Thank you very much
Peking University
April 10th, 1995

Please foreword this message to your freinds if you think they can help us, Thanks advanced!

e-mail: ca___@mccux0.mech.pku.edu.cn

End of messages
International Electronic Link Solves Medical Puzzle

DIPLOMATS FOSTERING the slow thaw in US-China relations might look to the medical community for inspiration.

Via the worldwide computer Internet and other means of communication, phys- sicians and other medical scientists from coast to coast in the United States and at least 17 other countries have helped their mainland China colleagues treat a university student with a challenging array of signs and symptoms.

The patient, Zhu Lingling, 21, a junior studying physical chemistry in Beijing, reportedly experienced abdominal pain and alopecia in December 1994 but returned to college in February after she responded to Chinese traditional therapy and nutritional support.

A month later, she was hospitalized again with a variety of central nervous system complaints and became comatose within 5 days.

After tests ruled out a number of tentative diagnoses and she did not respond significantly to treatment, students at Beijing University who had learned of the complicated case sent an electronic mail request for diagnostic and therapeu- tic assistance. This was relayed by several groups on the Internet.

Apparently the first to respond from the United States with the correct diagnosis was Stephen O. Cunnion, MD, PhD, MPH.

A US Navy captain who is an infectious disease epidemiologist in the Depart- ment of Preventive Medicine and Biostatistics, Uniformed Services University, Bethesda, Md, Cunnion diagnosed the problem as thallium poisoning.

In the next 4 weeks more than 80 other individuals and groups in medicine—out of some 2000 that eventually responded—supported Cunnion’s diagnosis. At least one physician in Beijing had suggested the thallium poisoning diagnosis, but it had not been put to the test there.

However, following the Internet response, the Chinese conducted tests that confirmed the thallium poisoning diagnosis. Physicians and other medical scientists in California, particularly the University of California, Los Angeles (UCLA), coordinated much of the succeeding effort with colleagues elsewhere.

Cunnion says that Chinese physicians now report the young woman has regained consciousness. The prognosis is encouraging, say many of the experts involved, but recovery—perhaps with only limited neurological sequelae—may be a long process.

The source of the poisoning has not been determined. One report indicates that Chinese authorities are looking into the possibility of criminal intent.

In the meantime, physicians and others involved in the evolving field of tele- medicine suggest that this experience offers some insight into its future po- tential.—by Phil Gunby

1750. JAMA, December 13, 1995—Vol 274, No. 22

Medical News & Perspectives
Wikipedia

From Wikipedia, the free encyclopedia

For Wikipedia's non-encyclopedic visitor introduction, see Wikipedia:About.

Wikipedia (pronounced /ˈwɪkədiə/ or /ˈwɜːkədiə/) is a free, web-based, collaborative, multilingual encyclopedia project supported by the non-profit Wikimedia Foundation. Its 18 million articles (over 3.6 million in English) have been written collaboratively by volunteers around the world, and almost all of its articles can be edited by anyone with access to the site.[3] As of May 2011, there were editions of Wikipedia in 281 languages. Wikipedia was launched in 2001 by Jimmy Wales and Larry Sanger[4] and has become the largest and most popular resource...
Revision history of Wikipedia

From Wikipedia, the free encyclopedia

View logs for this page

Browse history

From year (and earlier): From month (and earlier): all Tag filter: Deleted only Go

For any version listed below, click on its date to view it. For more help, see Help:Page history and Help:Edit summary.

External tools: Revision history statistics • Contributors • Revision history search • Number of watchers • Page view statistics

(cur) = difference from current version, (prev) = difference from preceding version, m = minor edit, ← = section edit, ← = automatic edit summary

(latest | earliest) View (newer 50 | older 50) (20 | 50 | 100 | 250 | 500)

Compare selected revisions

- (cur | prev) 21:34, 22 June 2011 Barek (talk | contribs) (145,219 bytes) (Undid revision 435705488 by Tfkjones (talk) rv - unsourced)
- (cur | prev) 21:20, 22 June 2011 Tfkjones (talk | contribs) m (145,463 bytes) (Satire)
- (cur | prev) 15:12, 21 June 2011 Rcsprinter123 (talk | contribs) (145,219 bytes) (bracket 'that anyone can edit' because there's also the slogan with just the ‘free encyclopedia’)
- (cur | prev) 15:11, 21 June 2011 Citation bot (talk | contribs) m (145,216 bytes) ([364]Add: chapter, isbn. Tweak: isbn, title.)
- (cur | prev) 19:41, 20 June 2011 Rich Farmbrough (talk | contribs) m (145,165 bytes) (Capitalise Wikipedia or other minor fixes. using AWB)
- (cur | prev) 01:16, 20 June 2011 Wxidea (talk | contribs) (145,168 bytes) (Reusing Wikipedia's content: Converted to bullets, added link to iOS apps)
- (cur | prev) 19:00, 19 June 2011 Pumagirl7 (talk | contribs) (144,670 bytes) (Satire)
- (cur | prev) 23:55, 17 June 2011 Rememberway (talk | contribs) m (144,357 bytes) (Defenses against undesirable edits)
- (cur | prev) 01:21, 17 June 2011 MarmadukePercy (talk | contribs) (144,344 bytes) (Undid revision 434685181 by Aidwniki (talk) rv rather
The Wisdom of Crowds and over 950,000 other books are available for Amazon Kindle – Amazon’s new wireless reading device. Learn more.
The Wisdom of Crowds

Customer Reviews

199 Reviews

Average Customer Review

5 star: 86 reviews
4 star: 53 reviews
3 star: 27 reviews
2 star: 20 reviews
1 star: 13 reviews

Search Customer Reviews

Share your thoughts with other customers

Only search this product's reviews

The most helpful favorable review

251 of 276 people found the following review helpful:

A Counter-Intuitive Notion
In 1906, Francis Galton, known for his work on statistics and heredity, came across a weight-judging contest at the West of England Fat Stock and Poultry Exhibition. This encounter was to challenge the foundations of his life's study.

The most helpful critical review

373 of 408 people found the following review helpful:

Accessible tome on behavioral economics and game theory.
What Other Items Do Customers Buy After Viewing This Item?

- **Blink: The Power of Thinking Without Thinking** by Malcolm Gladwell  Paperback
  $9.59
- **The Tipping Point: How Little Things Can Make a Big Difference** by Malcolm Gladwell  Paperback
  $9.52
- **Outliers: The Story of Success** by Malcolm Gladwell  Paperback
  $9.93
- **Here Comes Everybody: The Power of Organizing Without Organizations** by Clay Shirky  Paperback
  $10.88

> Explore similar items

Looking for "wisdom of crowds" Products?

Other customers suggested these items:

- **Infotopia: How Many Minds Produce Knowledge** by Cass R. Sunstein
  $19.19  (15)  Suggested by 1 customer
- **The Wealth of Networks: How Social Production Transforms Markets and Freedom** by Yochai Benkler
  $19.19  (15)  Suggested by 1 customer
- **The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business** by James Surowiecki
  $11.41  (199)  Suggested by 1 customer
- **The Cathedral & the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary** by Eric S. Raymond
  $11.41  (56)  Suggested by 1 customer
- **The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies** by Scott E. Page
  $11.41  (14)  Suggested by 1 customer
<table>
<thead>
<tr>
<th>Rank</th>
<th>Username</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>lubeckm</td>
<td>31M</td>
</tr>
<tr>
<td>2</td>
<td>greggig</td>
<td>5209K</td>
</tr>
<tr>
<td>3</td>
<td>Catwoman</td>
<td>4537K</td>
</tr>
<tr>
<td>4</td>
<td>pook</td>
<td>3283K</td>
</tr>
<tr>
<td>5</td>
<td>knowitall</td>
<td>2540K</td>
</tr>
</tbody>
</table>
KASPAROV AGAINST THE WORLD
THE STORY OF THE GREATEST ONLINE CHALLENGE
WORLD CHESS CHAMPION
GARRY KASPAROV
with Grandmaster Daniel King
MDS

Like professional baseball, we know a farm system is a better way to identify skilled investors.

For winning portfolios from the Marketocracy bench, find your best team at Marketocracy.com.

READ MORE
What is the IEM?

The IEM is an on-line futures market where contract payoffs are based on real-world events such as political outcomes, companies' earnings per share (EPS), and stock price returns. The market is operated by University of Iowa Henry B. Tippie College of Business faculty as an educational and research project. More...

Who can participate in the IEM?
Are the participants playing with real money?
Can markets predict the future?
Can I get historical data from the IEM?
How do I start trading?
I need more information about the IEM...

News

All

June 3, 2011
Dean Search Committee Named

February 23, 2011
Contest to Find Tippie MBA's Best Box Office Prognosticator

November 19, 2010
It's Too Early for the IEM to Predict the Next U.S. President

November 2, 2010
Brietz Searches for Politico-Economic Answers in 2010 Elections
Threadless is a community-based company that prints awesome designs created and chosen by you!
Welcome To InnoCentive
Where the World Innovates

Are you looking to solve problems and accelerate your innovation capability?

Drive Innovation »

Are you passionate about solving important problems that really matter?

Become A Solver »

Open Challenges

- **EDF – Nitrate Capture System**
  Deadline: 07/13/2011 | 193 active solvers | Referral award: $7,500 USD
  - **$7,500 USD**

- **Communication Platform to Connect Vulnerable Communities with Climate Change Solutions**
  Deadline: 07/12/2011 | 408 active solvers | Referral award: $1,000 USD
  - **$10,000 USD**

- **Eliminate Potholes – StreetBump for Boston!**
  Deadline: 07/29/2011 | 582 active solvers | Referral award: $2,500 USD
  - **$25,000 USD**

- **Educational GUI for Collaborative Problem Solving**
  Deadline: 07/05/2011 | 314 active solvers | Referral award: $2,000 USD
  - **$20,000 USD**

**NEWSFLASH**

2010 Top Solvers Honored, Solver Programs Launched
Congratulations to our 2010 Top Solvers! We’re also pleased to announce the InnoCentive Challenge Referral Program, which allows anyone to earn awards by referring Challenges to people who they believe can solve them. Also new is InnoCentive Anywhere, a cross-platform mobile app which allows smartphone users to access, follow, and share InnoCentive Challenges.

Learn More
Introduction

• Broad overview of collective intelligence, a framework for understanding it, and its various connections to machine learning
  - Animals to humans images/videos
  - CI mosaic (icons, then list of keywords)
  - CI, crowdsourcing, human computation
  - Overt vs covert

Collaborative creation

Collaborative decision making

Smartest in the crowd / contests

HC and micro-labor markets

Crowd mining

Roles of Machine Learning
Collective Intelligence

• “groups of individuals doing things collectively that seem intelligent”

“The Collective Intelligence Genome”, Malone, Laubacher and Dellarocas, MIT Management Review, April 1, 2010

• “How can people and computers be connected so that—collectively—they act more intelligently than any individuals, groups, or computers have ever done before?”

http://cci.mit.edu
Crowdsourcing

• “Crowdsourcing is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call.”

Crowdsourcing

• “We say that a system is a [crowdsourcing] system if it enlists a crowd of humans to help solve a problem defined by the system owners, and if in doing so, it addresses the following four fundamental challenges:
  – How to recruit and retain users?
  – What contributions can users make?
  – How to combine user contributions to solve the target problem?
  – How to evaluate users and their contributions?”

Human Computation

• “The idea behind digital computers may be explained by saying that these machines are intended to carry out any operations which could be done by a human computer.”

Human Computation

- Human computation is “a paradigm for utilizing human processing power to solve problems that computers cannot yet solve.”

Human Computation

• Human computation:
  – The problems fit the general paradigm of computation, and as such might someday be solvable by computers.
  – The human participation is directed by the computational system or process

Quinn and Bederson CHI 2011
Crowdsourcing vs Human Computation

• Whereas human computation replaces computers with humans, crowdsourcing replaces traditional human workers with members of the public.

Quinn and Bederson CHI 2011
Social Computing

• “applications and services that facilitate collective action and social interaction online with rich exchange of multimedia information and evolution of aggregate knowledge”

Collective Intelligence
+ Human
"Collective Intelligence"
- Human
Types of Collective Intelligence

• “Overt” vs “Covert”: Are human participants explicitly participating to achieve the collective outcomes, or is some form of mining of human activity achieving the collective outcomes
  – Overt: Amazon reviews, Wikipedia, AMT
  – Covert (“Crowd Mining”): Amazon recommendations, Google, Spam detection
Types of Collective Intelligence

• Overt
  – Collecting
  – Collaborative Creation
  – Smartest in the Crowd
  – Collaborative Decisions
  – Human Computation and Micro-Crowdsourcing

• Covert
  – Crowd Mining
Types of Collective Intelligence

• Collecting
Customer Reviews

The Wisdom of Crowds

199 Reviews

Average Customer Review

5 star: (86)
4 star: (53)
3 star: (27)
2 star: (20)
1 star: (13)

Search Customer Reviews

Share your thoughts with other customers

Create your own review

The most helpful favorable review

251 of 276 people found the following review helpful:

★★★★★★ A Counter-Intuitive Notion
In 1906, Francis Galton, known for his work on statistics and heredity, came across a weight-judging contest at the West of England Fat Stock and Poultry Exhibition. This encounter was to challenge the foundations of his life's study.

The most helpful critical review

373 of 408 people found the following review helpful:

★★★★☆ Accessible tome on behavioral economics and game theory.
Types of Collective Intelligence

• Collecting
  – Amazon reviews
Types of Collective Intelligence

- Collecting
  - Amazon reviews
  - YouTube
Types of Collective Intelligence

• Collecting
  – Amazon reviews
  – YouTube
  – Flickr
Hyatt Regency Bellevue
900 Bellevue Way NE, Bellevue, WA 98004-4272

TripAdvisor ranking
#6 of 26 hotels in Bellevue

172 Reviews

Check-in 7/8/2011  Check-out 7/10/2011  Adults 1

Show Prices

Expedia.com  ORBITZ.com  Priceline.com  Travelocity
Hotels.com  Hyatt.com  Venere.com

Recently viewed
Silver Cloud Inn Bellevue
La Residence Suite Hotel
Franklin Cafe Cape Ann
The Red Skiff Seafood & Grille
Seaport Grille

Save  E-mail  Clear

Browse nearby
Hotels (26)  Restaurants (239)  Things to Do (19)

Map of Hyatt Regency Bellevue

See professional photos
See traveler photos (22)
“Great business hotel”

Reviewed June 15, 2011

Modern, spacious rooms with fairly typical Hyatt decor. Excellent, really top-notch gym with lots of equipment and weights: I loved this.
Types of Collective Intelligence

- **Collecting**: Assembling large collections of online materials
  - Amazon reviews
  - YouTube
  - Flickr
  - TripAdvisor
Types of Collective Intelligence

• Collecting: Assembling large collections of online materials
  – Amazon reviews
  – YouTube
  – Flickr
  – TripAdvisor
  – Ten Thousand Cents
Types of Collective Intelligence

- **Collaborative Creation**: Jointly crafting on-line materials that are more holistic
Types of Collective Intelligence

• **Collaborative Creation**: Jointly crafting on-line materials that are more holistic
  – Wikipedia
Types of Collective Intelligence

- **Collaborative Creation**: Jointly crafting on-line materials that are more holistic
  - Wikipedia
  - Open Source Software
A NEW PROOF OF THE DENSITY HALES-JEWETT THEOREM

D. H. J. POLYMATH

ABSTRACT. The Hales–Jewett theorem asserts that for every $r$ and every $k$ there exists $n$ such that every $r$-colouring of the $n$-dimensional grid $\{1, \ldots, k\}^n$ contains a combinatorial line. This result is a generalization of van der Waerden’s theorem, and it is one of the fundamental results of Ramsey theory. The theorem of van der Waerden has a famous density version, conjectured by Erdős and Turán in 1936, proved by Szemerédi in 1975, and given a different proof by Furstenberg in 1977. The Hales–Jewett theorem has a density version as well, proved by Furstenberg and Katznelson in 1991 by means of a significant extension of the ergodic techniques that had been pioneered by Furstenberg in his proof of Szemerédi’s theorem. In this paper, we give the first elementary proof of the theorem of Furstenberg and Katznelson, and the first to provide a quantitative bound on how large $n$ needs to be. In particular, we show that a subset of $\{1, 2, 3\}^n$ of density $\delta$ contains a combinatorial line if $n$ is at least as big as a tower of 2s of height $O(1/\delta^2)$. Our proof is surprisingly simple: indeed, it gives arguably the simplest known proof of Szemerédi’s theorem.

1. Introduction

1.1. Statement of our main result. The purpose of this paper is to give the first elementary proof of the density Hales–Jewett theorem. This theorem, first proved by Furstenberg and Katznelson [FK89, FK91], has the same relation to the Hales–Jewett theorem as Szemerédi’s theorem [Sze75] has to van der Waerden’s theorem [vdW27]. Before we go any further, let us state all four theorems. We shall use the notation $[k]$ to stand for the set $\{1, \ldots, k\}$.
Types of Collective Intelligence

• Collaborative Creation: Jointly crafting on-line materials that are more holistic
  – Wikipedia
  – Open Source Software
  – Gowers’ Polymath Project
Types of Collective Intelligence

- **Smartest in the Crowd**: People compete for some prize or recognition
  - Innocentive
  - Goldcorp Challenge
  - Threadless
  - Netflix Challenge
Types of Collective Intelligence

• Collaborative Decisions: People’s efforts serve to make selections out of some list of options
  – Iowa Electronic Markets
  – Ebbsfleet United
  – Kasparov against the World
Types of Collective Intelligence

• Human Computation and Micro-Crowdsourcing: Individuals perform numerous small tasks that are hard for computers and that collectively solve larger, more difficult problems
Welcome to Galaxy Zoo, where you can help astronomers explore the Universe

Galaxy Zoo: Hubble uses gorgeous imagery of hundreds of thousands of galaxies drawn from NASA's Hubble Space Telescope archive. To
Types of Collective Intelligence

• **Human Computation and Micro-Crowdsourcing:** Individuals perform numerous small tasks that are hard for computers and that collectively solve larger, more difficult problems
  – GalaxyZoo
Types of Collective Intelligence

- **Human Computation and Micro-Crowdsourcing**: Individuals perform numerous small tasks that are hard for computers and that collectively solve larger, more difficult problems
  - GalaxyZoo
  - ReCAPTCHA
SAILOR MISSING SINCE 1/28/07
Please contact the United States Coast Guard with any information.

Wired Article   NY Times Article   Ongoing Effort   I'd Like to Help!   Print a MISSING Poster

Announcement: Satellite Image Examination Done! We've examined more than 560,000 images from 3 satellites, covering nearly 3,500 square miles of ocean! We currently do not need help here, but are looking for help elsewhere.

Sailboat: TENACIOUS

40 ft C&C Sailboat      Sail # 31869
Red Hull, Black Mast, Silver/White Side Stripes

Sailor: JIM GRAY

6'3”, 190 lbs. Gray Hair, White Beard
63 yrs old      Brown Eyes, Thick Eyebrows

Jim may have been wearing a blue or white shirt, cream colored crewneck sweater, and black rain jacket over levi jeans. Sailed from the Golden Gate Bridge in San Francisco, California, to the Farallon Islands (25 miles west of San
Types of Collective Intelligence

• **Human Computation and Micro-Crowdsourcing:** Individuals perform numerous small tasks that are hard for computers and that collectively solve larger, more difficult problems
  - GalaxyZoo
  - ReCAPTCHA
  - Search for Jim Gray
Types of Collective Intelligence

• **Human Computation and Micro-Crowdsourcing:** Individuals perform numerous small tasks that are hard for computers and that collectively solve larger, more difficult problems
  – GalaxyZoo
  – ReCAPTCHA
  – Search for Jim Gray
  – “Games with a Purpose”
    • ESP Game
Predicting protein structures with a multiplayer online game

Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović & Foldit players

Affiliations | Contributions | Corresponding authors

Nature 466, 756–760 (05 August 2010) | doi:10.1038/nature09304
Received 22 January 2010 | Accepted 30 June 2010

People exert large amounts of problem-solving effort playing computer games. Simple image- and text-recognition tasks have been successfully ‘crowd-sourced’ through games1, 2, 3, but it is not clear if more complex scientific problems can be solved with human-directed computing. Protein structure prediction is one such problem: locating the biologically relevant native conformation of a protein is a formidable computational challenge given
Types of Collective Intelligence

• **Human Computation and Micro-Crowdsourcing:** Individuals perform numerous small tasks that are hard for computers and that collectively solve larger, more difficult problems
  – GalaxyZoo
  – ReCAPTCHA
  – Search for Jim Gray
  – “Games with a Purpose“
    • ESP Game
    • fold.it
Types of Collective Intelligence

• **Human Computation and Micro-Crowdsourcing:** Individuals perform numerous small tasks that are hard for computers and that collectively solve larger, more difficult problems
  – GalaxyZoo
  – ReCAPTCHA
  – Search for Jim Gray
  – “Games with a Purpose“
    • ESP Game
    • fold.it
  – Amazon Mechanical Turk
Cheap and Fast — But is it Good?
Evaluating Non-Expert Annotations for Natural Language Tasks

Rion Snow† Brendan O’Connor‡ Daniel Jurafsky§ Andrew Y. Ng†

†Computer Science Dept.
Stanford University
Stanford, CA 94305
{rion, ang}@cs.stanford.edu

‡Dolores Labs, Inc.
832 Capp St.
San Francisco, CA 94110
brendano@doloreslabs.com

§Linguistics Dept.
Stanford University
Stanford, CA 94305
jurafsky@stanford.edu

Abstract

Human linguistic annotation is crucial for many natural language processing tasks but can be expensive and time-consuming. We explore the use of Amazon’s Mechanical Turk system, a significantly cheaper and faster method for collecting annotations from a broad base of paid non-expert contributors over the Web. We investigate five tasks: affect recognition, word similarity, recognizing textual entailment, event temporal ordering, and word sense disambiguation. For all five, we show high agreement between Mechanical Turk non-expert annotations and existing gold standard labels provided by expert labelers. For the task of affect recognition, we also and financial cost. Since the performance of many natural language processing tasks is limited by the amount and quality of data available to them (Banko and Brill, 2001), one promising alternative for some tasks is the collection of non-expert annotations.

In this work we explore the use of Amazon Mechanical Turk (AMT) to determine whether non-expert labelers can provide reliable natural language annotations. We chose five natural language understanding tasks that we felt would be sufficiently natural and learnable for non-experts, and for which we had gold standard labels from expert labelers, as well as (in some cases) expert labeler agree-
Utility data annotation with Amazon Mechanical Turk

Alexander Sorokin, David Forsyth
University of Illinois at Urbana-Champaign
201 N Goodwin
Urbana, IL 61820
{xorokin2, daf}@uiuc.edu

Abstract

We show how to outsource data annotation to Amazon Mechanical Turk. Doing so has produced annotations in quite large numbers relatively cheaply. The quality is good, and can be checked and controlled. Annotations are produced quickly. We describe results for several different annotation problems. We describe some strategies for determining when the task is well specified and properly priced.

1. Introduction

Big annotated image datasets now play an important role in Computer Vision research. Many of them were built in-house ([18, 11, 12, 3, 13, 5] and many others). This consumes significant amounts of highly skilled labor, requires much management work, is expensive and creates a perception that annotation is difficult. Another successful strategy is to make the annotation process completely public.

<table>
<thead>
<tr>
<th>Exp</th>
<th>Task</th>
<th>img</th>
<th>labels</th>
<th>cost</th>
<th>time</th>
<th>effective pay/hr</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>170</td>
<td>510</td>
<td>$8</td>
<td>750m</td>
<td>$0.76</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>170</td>
<td>510</td>
<td>$8</td>
<td>380m</td>
<td>$0.77</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>305</td>
<td>915</td>
<td>$14</td>
<td>950m</td>
<td>$0.41</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>305</td>
<td>915</td>
<td>$14</td>
<td>150m</td>
<td>$1.07</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>337</td>
<td>1011</td>
<td>$15</td>
<td>170m</td>
<td>$0.9</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>982</td>
<td>3861</td>
<td>$59</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Collected data. In our five experiments we have collected 3861 labels for 982 distinct images for only US $59. In experiments 4 and 5 the throughput exceeds 300 annotations per hour even at low ($1/hour) hourly rate. We expect further increase in throughput as we increase the pay to effective market rate.

a researcher are: (1) define an annotation protocol and (2) determine what data needs to be annotated.

The annotation protocol should be implemented within an IFRAME of a web browser. We call the implementation
USING THE AMAZON MECHANICAL TURK
FOR TRANSCRIPTION OF SPOKEN LANGUAGE

Matthew Marge, Satrajit Banerjee, and Alexander I. Rudnicky

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
{mrmar, banerje, air}@cs.cmu.edu

ABSTRACT

We investigate whether Amazon’s Mechanical Turk (MTurk) service can be used as a reliable method for transcription of spoken language data. Utterances with varying speaker demographics (native and non-native English, male and female) were posted on the MTurk marketplace together with standard transcription guidelines. Transcriptions were compared against transcriptions carefully prepared in-house through conventional (manual) means. We found that transcriptions from MTurk workers were generally quite accurate. Further, when transcripts for the same utterance produced by multiple workers were combined using the ROVER voting scheme, the accuracy of the combined transcript rivaled that observed for conventional transcription methods. We also found that accuracy is not particularly sensitive to payment amount, implying that high quality results can be obtained at a fraction of the cost and turnaround time of conventional methods.

Index Terms— crowd sourcing, speech transcription

("requesters"). The requester determines when work is satisfactory. The tasks targeted by the system are ones that are simple for humans to perform but are challenging for computers (e.g., determining if a person is facing to the right or left in a picture). Speech transcription fits well into this framework. In this paper, we investigate whether MTurk can be used to perform transcription according to the conventions and quality level expected by the speech research community.

MTurk has been previously used by others to transcribe speech. For example, [1] and [2] report “near-expert accuracy” when using MTurk to correct the output of an automatic speech recognizer. MTurk has been used for other natural language annotation tasks. For example, [3] used MTurk to carry out several different annotation tasks (such as word sense disambiguation) and found strong agreement with gold standard annotations. [4] used MTurk to evaluate machine translation output. [5] evaluated MTurk as a venue
Creating Speech and Language Data With Amazon’s Mechanical Turk

Chris Callison-Burch and Mark Dredze
Human Language Technology Center of Excellence
& Center for Language and Speech Processing
Johns Hopkins University
ccb, mdredze@cs.jhu.edu

Abstract

In this paper we give an introduction to using Amazon’s Mechanical Turk crowdsourcing platform for the purpose of collecting data for human language technologies. We survey the papers published in the NAACL-2010 Workshop. 24 researchers participated in the workshop’s shared task to create data for speech and language applications with $100.

1 Introduction

This paper gives an overview of the NAACL-2010 Turk to spend on an annotation task of their choosing. They were required to write a short paper describing their experience, and to distribute the data that they created. They were encouraged to address the following questions: How did you convey the task in terms that were simple enough for non-experts to understand? Were non-experts as good as experts? What did you do to ensure quality? How quickly did the data get annotated? What is the cost per label? Researchers submitted a 1 page proposal to the workshop organizers that described their intended experiments and expected outcomes. The organizers selected proposals based on merit, and awarded $100 credits that were generously provided.
Can Crowds Build Parallel Corpora for Machine Translation Systems?

Vamshi Ambati and Stephan Vogel
{vamshi,vogel}@cs.cmu.edu
Language Technologies Institute, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213, USA

Abstract

Corpus based approaches to machine translation (MT) rely on the availability of parallel corpora. In this paper we explore the effectiveness of Mechanical Turk for creating parallel corpora. We explore the task of sentence translation, both into and out of a language. We also perform preliminary experiments for the task of phrase translation, where ambiguous phrases are provided to the turker for translation in isolation and in the context of the sentence it originated from.

1 Introduction

machine translation (Ambati et al., 2010; Callison-Burch, 2009), but focuses primarily on the setup and design of translation tasks on MTurk with varying granularity levels, both at sentence- and phrase-level translation.

2 Language Landscape on MTurk

We first conduct a pilot study by posting 25 sentences each from a variety of language pairs and probing to see the reception on MTurk. Language-pair selection was based on number of speakers in the language and Internet presence of the population. Languages like Spanish, Chinese, English,
Crowdsourcing User Studies With Mechanical Turk

Aniket Kittur, Ed H. Chi, Bongwon Suh
Palo Alto Research Center
3333 Coyote Hill Road, Palo Alto, CA 94304 USA
{nkittur, echi, suh}@parc.com

ABSTRACT
User studies are important for many aspects of the design process and involve techniques ranging from informal surveys to rigorous laboratory studies. However, the costs involved in engaging users often requires practitioners to trade off between sample size, time requirements, and monetary costs. Micro-task markets, such as Amazon’s Mechanical Turk, offer a potential paradigm for engaging a large number of users for low time and monetary costs. Here we investigate the utility of a micro-task market for collecting user measurements, and discuss design considerations for developing remote micro user evaluation tasks. Although micro-task markets have great potential for rapidly collecting user measurements at low costs, we found that special care is needed in formulating tasks in order to harness the capabilities of the approach.

Author Keywords
Remote user study, Mechanical Turk, micro task, Wikipedia.

ACM Classification Keywords

others. Thus it is often not possible to acquire user input that is both low-cost and timely enough to impact development. The high costs of sampling additional users lead practitioners to trade off the number of participants with monetary and time costs [5].

Collecting input from only a small set of participants is problematic in many design situations. In usability testing, many issues and errors (even large ones) are not easily caught with a small number of participants [5]. In both prototyping and system validation, small samples often lead to a lack of statistical reliability, making it difficult to determine whether one approach is more effective than another. The lack of statistical rigor associated with small sample sizes is also problematic for both experimental and observational research.

These factors have led to new ways for practitioners to collect input from users on the Web, including tools for user surveys (e.g., surveymonkey.com, vividence.com), online experiments [3], and remote usability testing [2]. Such tools expand the potential user pool to anyone connected to the
Crowdsourcing Preference Judgments for Evaluation of Music Similarity Tasks

Julián Urbano, Jorge Morato, Mónica Marrero and Diego Martín
University Carlos III of Madrid
Department of Computer Science
Leganés, Madrid, Spain
{jurbano, jmorato, mmarrero, dm1andres}@inf.uc3m.es

ABSTRACT
Music similarity tasks, where musical pieces similar to a query should be retrieved, are quite troublesome to evaluate. Ground truths based on partially ordered lists were developed to cope with problems regarding relevance judgment, but they require such manpower to generate that the official MIREX evaluations had to turn over more affordable alternatives. However, in house evaluations keep using these partially ordered lists because they are still more suitable for similarity tasks. In this paper we propose a cheaper alternative to generate these lists by using crowdsourcing to gather music preference judgments. We show that our method produces lists very similar to the original ones, while dealing with some defects of the original methodology. With this study, we show that crowdsourcing is a perfectly viable alternative to evaluate music systems without the need for experts.

Categories and Subject Descriptors
H.5.1 [Multimedia Information Systems]: Evaluation/methodology; H.3.3 [Information Search and Retrieval]; H.3.4 [Systems and Software]: Performance evaluation (efficiency and effectiveness).

Keywords
Crowdsourcing, relevance judgment, music information retrieval.

1. INTRODUCTION

the case of the Symbolic Melodic Similarity (SMS) and Audio Music Similarity (AMS) tasks, as defined in MIREX, in which systems are asked to retrieve a ranked list of musical pieces deemed similar to some piece of music acting as query. In particular, it is unclear how to assess the relevance of a document for a given query.

Ground truths are traditionally based on a fixed scale of relevance with levels such as “relevant” and “not relevant”. However, several studies indicate that relevance is continuous for information needs involving music similarity [4][5][6]. Single melodic changes such as moving a note up or down in pitch, or extending or shortening its duration, are not perceived to change the overall melody. However, the relationship with the original melody is gradually weaker as more changes are applied to it. There are no common criteria to split the degree of relevance into different levels, so assessments based on a fixed scale do not seem suitable as it would be difficult to draw the line between levels.

Major advancements in this matter were achieved by Typke et al. by the beginning of 2005. They developed a methodology to create ground truths where the relevance of a document does not belong to any prefixed scale, but it is rather implied by its relative position in a partially ordered list [5]. These lists are ordered so that the earlier a group appears in the list, the more relevant its documents are (see Figure 2). That way, the ideal retrieval
Can we get rid of TREC assessors?
Using Mechanical Turk for relevance assessment

Omar Alonso
A9.com
Palo Alto, CA (USA)
oralonso@gmail.com

Stefano Mizzaro
Dept. of Mathematics and Computer Science
University of Udine
Udine (Italy)
mizzaro@dimi.uniud.it

ABSTRACT
Recently, Amazon Mechanical Turk has gained a lot of attention as a tool for conducting different kinds of relevance evaluations. In this paper we show a series of experiments on TREC data, evaluate the outcome, and discuss the results. Our position, supported by these preliminary experimental results, is that crowdsourcing is a viable alternative for relevance assessment.

Categories and Subject Descriptors
H.3.4 [Information Storage and Retrieval]: Systems and software — performance evaluation

General Terms
Measurement, performance, experimentation

Keywords
IR evaluation, relevance, relevance assessment, user study

1. INTRODUCTION AND MOTIVATIONS
One issue in current TREC-like test collection initiatives is the cost related to relevance assessment: assessing requires resources (that cost time and even money) and does not scale up. Indeed, in recent years, there has been some trend on trying to save assessment resources: there is a vast body of literature on reducing the number of documents pooled and/or judged, and, more recently, on reducing the number of assessed topics [4] as well. Also, test collections are sometimes built in-house [4], and assessment effort is obviously a

<table>
<thead>
<tr>
<th>Table 1: The five experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1 Graded relevance on a 4 point scale (3 = excellent, 2 = good, 1 = fair, 0 = not relevant) following closely TREC-7 guidelines. We summarized the main points from the TREC assessment guidelines as starting point.</td>
</tr>
<tr>
<td>E2 Graded relevance with modified instructions. Changes on the instructions, use more layman English (not so expert). We also included an input form in the task so turkers can provide feedback.</td>
</tr>
<tr>
<td>E3 Graded relevance with modified instructions II. Modified instructions using colors and examples of relevant content. Also included more documents in the test.</td>
</tr>
<tr>
<td>E4 Binary relevance without qualification test. Maintained same instructions but changed the answers to binary (1 = relevant and 0 = not relevant). Modified the feedback input to an optional entry for justifying answers. Passing grade was 80% of correct answers.</td>
</tr>
<tr>
<td>E5 Binary relevance with qualification test. Same as previous experiment but with a lower passing grade for the qualification test to 60%.</td>
</tr>
</tbody>
</table>

is known as the requester. A person who wants to sign up to perform work is described in the system as a turker. The unit of work to be performed is called a HIT (Human Intelligence Task). Each HIT has an associated payment and an allotted completion time. It is possible to control the quality of the work by using qualification tests. MTurk has already been used in some relevance related research [1][2][3], with good success.

Therefore, our research question can be framed as: “Is it possible to replace TREC-like relevance assessors with Mechanical turkers?” We think the answer is “Yes — at least
Crowdsourcing a News Query Classification Dataset

Richard M. C. McCreadie  
Department of Computing Science  
University of Glasgow  
Glasgow, G12 8QQ  
richardm@dcs.gla.ac.uk

Craig Macdonald  
Department of Computing Science  
University of Glasgow  
Glasgow, G12 8QQ  
craigm@dcs.gla.ac.uk

Iadh Ounis  
Department of Computing Science  
University of Glasgow  
Glasgow, G12 8QQ  
onis@dcs.gla.ac.uk

ABSTRACT

Web search engines are well known for aggregating news vertical content into their result rankings in response to queries classified as news-related. However, no dataset currently exists upon which approaches to news query classification can be evaluated and compared. This paper studies the generation and validation of a news query classification dataset comprised of labels crowdsourced from Amazon’s Mechanical Turk and details insights gained. Notably, our study focuses around two challenges when crowdsourcing news query classification labels: 1) how to overcome our workers’ lack of information about the news stories from the time of each query and 2) how to ensure the resulting labels are of high enough quality to make the dataset useful. We empirically show that a worker’s lack of information about news stories can be addressed through the integration of news-related content into the labelling interface and that this improves the quality of the resulting labels. Overall, we find that crowdsourcing is suitable for building a news query classification dataset.

General Terms: Performance, Experimentation

Keywords: News Query Classification, Crowdsourcing, Vertical Search

1. INTRODUCTION

General-purpose Web search engines are well known for integrating focused news vertical content into their search rankings dataset. We propose multiple interfaces for crowdsourced query labelling and evaluate these interfaces empirically in terms of the quality of the resulting labels on a small representative sample of user queries from a Web search engine query log. Later, we use the best performing of these interfaces to generate our final news query classification dataset comprised of a larger query sample from the same log. We report the quality of our resulting news query classification dataset in terms of inter-worker labelling agreement and accuracy with regard to labels created separately by the authors. Moreover, we further investigate its quality in the form of an additional agreement study, in which crowdsourcing is leveraged for quality assurance.

Notably, one of the most interesting aspects of news query classification labelling is the temporal nature of news-related queries [16]. In particular, a query should only be labelled as news-related if there was a relevant noteworthy story in the news around the time each query was made. However, the query log we employ dates back to 2006 [7], hence there is no guarantee that our workers will remember what the major news stories were from the time of each query. In this work, we empirically investigate the effect that this has on labelling quality. Moreover, we propose the integration of news headlines, article summaries and Web search interface seen by the workers to address this problem.

The main contributions of this paper are four-fold. Firstly, we examine the suitability of crowdsourcing for the creation of a news query classification dataset; secondly we both propose and evaluate
How Well Do Line Drawings Depict Shape?

Forrester Cole¹  Kevin Sanik²  Doug DeCarlo²
Adam Finkelstein¹  Thomas Funkhouser¹  Szymon Rusinkiewicz¹,³  Manish Singh²
¹Princeton University  ²Rutgers University  ³Adobe Systems

Figure 1: Gauge figure results. In this study, people were shown one of six different renderings of a shape: (a) a shaded image, (b) a line drawing made from the shaded image by a person, (c) contours, (d) apparent ridges, and (shown in Figure 7) ridges/valleys and suggestive contours. Overlaid are representative “gauges” (discs revealing the surface normal) oriented on the images by people in the study, colored by how far they deviate from the ground truth.

Abstract

This paper investigates the ability of sparse line drawings to depict 3D shape. We perform a study in which people are shown an image of one of twelve 3D objects depicted with one of six styles and asked to orient a gauge to coincide with the surface normal at many positions on the object’s surface. The normal estimates are compared with each other and with ground truth data provided by a registered 3D surface model to analyze accuracy and precision. The paper describes the design decisions made in collecting a large data set (275,000 gauge measurements) and provides analysis to answer questions about how well people interpret shapes from drawings. Our findings suggest that people interpret certain shapes almost as well from a line drawing as from a shaded image, that current some line drawings are more effective than others at doing this; and that some shapes are difficult to draw effectively. However, there is little scientific evidence in the literature for these observations. Moreover, while a recent thread of the computer graphics literature devoted to automatic algorithms for line drawings is flourishing, to date researchers have had no objective way to evaluate the effectiveness of such algorithms in depicting shape.

In this paper, we investigate how accurately people interpret shapes depicted by line drawings. At first, this goal seems difficult to achieve. Aside from asking sculptors to craft the shape they see, how can we know what shape is in a person’s mind? Koenderink et al. [1992] have proposed several strategies for experimentally measuring perceived geometry, based on collecting the results of many simple questions. This paper describes such
Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design

Jeffrey Heer and Michael Bostock
Computer Science Department
Stanford University
{jheer, mbostock}@cs.stanford.edu

ABSTRACT
Understanding perception is critical to effective visualization design. With its low cost and scalability, crowdsourcing presents an attractive option for evaluating the large design space of visualizations; however, it first requires validation. In this paper, we assess the viability of Amazon’s Mechanical Turk as a platform for graphical perception experiments. We replicate previous studies of spatial encoding and luminance contrast and compare our results. We also conduct new experiments on rectangular area perception (as in treemaps or cartograms) and on chart size and gridline spacing. Our results demonstrate that crowdsourced perception experiments are viable and contribute new insights for visualization design. Lastly, we report cost and performance data from our experiments and distill recommendations for the design of crowdsourced studies.

ACM Classification: H5.2 [Information interfaces and presentation]: User Interfaces—Evaluation/Methodology
General Terms: Experimentation, Human Factors.
Keywords: Information visualization, graphical perception.
Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers

Victor S. Sheng  
ssheng@stern.nyu.edu

Foster Provost  
fprovost@stern.nyu.edu

Panagiotis G. Ipeirotis  
panos@stern.nyu.edu

Department of Information, Operations, and Management Sciences  
Leonard N. Stern School of Business, New York University

ABSTRACT

This paper addresses the repeated acquisition of labels for data items when the labeling is imperfect. We examine the improvement (or lack thereof) in data quality via repeated labeling, and focus especially on the improvement of training labels for supervised induction. With the outsourcing of small tasks becoming easier, for example via Rent-A-Coder or Amazon’s Mechanical Turk, it often is possible to obtain less-than-expert labeling at low cost. With low-cost labeling, preparing the unlabeled part of the data can become considerably more expensive than labeling. We present repeated-labeling strategies of increasing complexity, and show several main results. (i) Repeated-labeling can improve label quality and model quality, but not always. (ii) When labels are noisy, repeated labeling can be preferable to single labeling even in the traditional setting where labels are not particularly cheap. (iii) As soon as the cost of processing the unlabeled data is not free, the simple strategy of labeling everything multiple times has considerable advantage. (iv) Repeatedly labeling a carefully chosen set of points is generally preferable, and we present a robust technique that combines different notions of uncertainty to select data points for which quality should be improved. The bottom line: the results show clearly that repeated-labeling can improve data quality and model quality.

1. INTRODUCTION

There are various costs associated with the preprocessing stage of the KDD process, including costs of acquiring features, formulating data, cleaning data, obtaining expert labeling of data, and so on [31, 32]. For example, in order to build a model to recognize whether two products described on two web pages are the same, one must extract the product information from the pages, formulate features for comparing the two along relevant dimensions, and label product pairs as identical or not; this process involves costly manual intervention at several points. To build a model that recognizes whether an image contains an object of interest, one first needs to take pictures in appropriate contexts, sometimes at substantial cost.

This paper focuses on problems where it is possible to obtain certain (noisy) data values (“labels”) relatively cheaply, from multiple sources (“labelers”). A main focus of this paper is the use of these values as training labels for supervised modeling. For our two examples above, once we have constructed the unlabeled example, for relatively low cost one can obtain non-expert opinions on whether an image contains an object of interest, or whether an image contains a particular product. These cheap labels may be noisy due to lack of expertise, dedication, interest, or other factors. Our ability to perform any expert labeling cheaply and easily is facilitated.
Active Learning and Crowd-Sourcing for Machine Translation

Vamshi Ambati, Stephan Vogel, Jaime Carbonell

Language Technologies Institute
Carnegie Mellon University, Pittsburgh, PA 15213, USA
{vamshi,vogel,jgc}@cs.cmu.edu

Abstract

In recent years, corpus based approaches to machine translation have become predominant, with Statistical Machine Translation (SMT) being the most actively progressing area. Success of these approaches depends on the availability of parallel corpora. In this paper we propose Active Crowd Translation (ACT), a new paradigm where active learning and crowd-sourcing come together to enable automatic translation for low-resource language pairs. Active learning aims at reducing cost of label acquisition by prioritizing the most informative data for annotation, while crowd-sourcing reduces cost by using the power of the crowds to make do for the lack of expensive language experts. We experiment and compare our active learning strategies with strong baselines and see significant improvements in translation quality. Similarly, our experiments with crowd-sourcing on Mechanical Turk have shown that it is possible to create parallel corpora using non-experts and with sufficient quality assurance, a translation system that is trained using this corpus approaches expert quality.

1. Introduction

Corpus based approaches to automatic translation like Example Based and Statistical Machine Translation systems use large amounts of parallel data created by humans to train mathematical models for automatic language translation (Koehn et al., 2003). Large scale parallel data generation for new language pairs requires intensive human effort and availability of experts. It becomes immensely difficult and costly to provide Statistical Machine Translation (SMT) systems for most languages due to the paucity of expert translators to provide parallel data. Even if active learning approaches help us identify sentences, which if translated have the potential to provide maximal improvement to an existing system. Crowd-sourcing techniques, on the other hand help us reach a significant number of translators at very low costs. This is very apt in a minority language scenario, where cost plays a major role. This paper addresses the following contributions:

- We propose and implement an end-to-end human-in-the-loop translation system framework called Active Crowd Translation, that combines online, non-expert, human translators with an automatic MT system.
The Multidimensional Wisdom of Crowds

Peter Welinder\textsuperscript{1}   Steve Branson\textsuperscript{2}   Serge Belongie\textsuperscript{2}   Pietro Perona\textsuperscript{1}
\textsuperscript{1} California Institute of Technology, \textsuperscript{2} University of California, San Diego
\{welinder,perona\}@caltech.edu   \{sbranson,sjb\}@cs.ucsd.edu

Abstract

Distributing labeling tasks among hundreds or thousands of annotators is an increasingly important method for annotating large datasets. We present a method for estimating the underlying value (e.g. the class) of each image from (noisy) annotations provided by multiple annotators. Our method is based on a model of the image formation and annotation process. Each image has different characteristics that are represented in an abstract Euclidean space. Each annotator is modeled as a multidimensional entity with variables representing competence, expertise and bias. This allows the model to discover and represent groups of annotators that have different sets of skills and knowledge, as well as groups of images that differ qualitatively. We find that our model predicts ground truth labels on both synthetic and real data more accurately than state of the art methods. Experiments also show that our model, starting from a set of binary labels, may discover rich information, such as different “schools of thought” amongst the annotators, and can group together images belonging to separate categories.

1 Introduction
Why Label when you can Search? Alternatives to Active Learning for Applying Human Resources to Build Classification Models Under Extreme Class Imbalance

Josh Attenberg
Polytechnic Institute of NYU
Brooklyn, NY
josh@cis.poly.edu

Foster Provost
NYU Stern School of Business
New York, NY
fprovost@stern.nyu.edu

ABSTRACT
This paper analyses alternative techniques for deploying low-cost human resources for data acquisition for classifier induction in domains exhibiting extreme class imbalance—where traditional labeling strategies, such as active learning, can be ineffective. Consider the problem of building classifiers to help brands control the content adjacent to their on-line advertisements. Although frequent enough to worry advertisers, objectionable categories are rare in the distribution of impressions encountered by most on-line advertisers—so rare that traditional sampling techniques do not find enough positive examples to train effective models. An alternative way to deploy human resources for training-data acquisition is to have them “guide” the learning by searching explicitly for training examples of each class. We show that under extreme skew, even basic techniques for guided learning completely dominate smart (active) strategies for applying human resources to select cases for labeling. Therefore, it is critical to consider the relative cost of search versus labeling, and we demonstrate the tradeoffs for different relative costs. We show that in cost/skew settings where the choice

1. INTRODUCTION
This paper concerns the interaction of humans in the data acquisition phase of the process of building classification models from data. Consider the following example data mining application: classifying web pages for the purpose of safe advertising. Advertisers and advertising networks (hereafter, advertisers) would like a rating system that estimates whether a web page or web site displays certain objectionable content. With such a system, advertisers can control the destination of their ads, advertising only on those pages deemed unlikely to display such unacceptable content (depending on the advertiser, objectionable categories include: adult content, kids content, hate speech, malware, etc.). Evaluating each potential advertising opportunity involves classifying the web page with respect to these objectionable categories. The classification system can take into account various evidence, including the URL, the page text, anchor text, DMOZ categories, position in the network, etc. [3]. For this paper, we will consider only the textual html source for each page, but the ideas generalize to any type of
CrowdFlow: Integrating Machine Learning with Mechanical Turk for Speed-Cost-Quality Flexibility

Alexander J. Quinn¹, Benjamin B. Bederson¹,²,³, Tom Yeh³, Jimmy Lin¹,²
University of Maryland
Human-Computer Interaction Lab
Department of Computer Science¹ || iSchool² || Institute for Advanced Computer Studies³
College Park, MD 20742 USA
aq@cs.umd.edu, bederson@cs.umd.edu, tomyeh@umiacs.umd.edu, jimmylin@umd.edu

Abstract
Humans and machines have competing strengths for tasks such as natural language processing and image understanding. Whereas humans do these things naturally with potentially high accuracy, machines offer greater speed and flexibility. CrowdFlow is our toolkit for a model for blending the two in order to attain tighter control over the inherent tradeoffs in speed, cost and quality. With CrowdFlow, humans and machines work together to do a set of tasks at a user-specified point in the tradeoff space. They work symbiotically, with the humans providing training data to the machine while the machine provides first cut results to the humans to save effort in cases where the machine’s answer was already correct. The CrowdFlow toolkit can be considered as a generalization of our other domain-specific efforts aimed at enabling cloud computing services using a variety of computational resources to achieve various tradeoff points.

1. Introduction
There is a large set of problems that can be solved either by human computation or machine learning. These include participants in collaboration with machine translation systems to translate books - resulting in translations which are between pure machine and pure expert (bilingual) human solutions [2][9]. Although MonoTrans is not a GWAP, it is similar in that it yields relatively high quality and low cost, at the expense of time. Naïve uses of AMT provide faster turnaround time (and a greater range of potential applications) at the expense of cost and quality.

The principal goal of this research is to discover strategies that combine the strengths of humans and machines, in order to move past these two rigid extremes. We would like to give developers the flexibility to choose any point in the speed-cost-quality spectrum, and automatically combine human and machine resources to achieve the desired balance. To do so in a general way is an ambitious goal.

We are developing a framework for blending capabilities of humans and machines in order to enable much greater flexibility with respect to speed-cost-quality. The strategy blends the flexibility of Amazon Mechanical Turk with machine learning (or other automated methods).
Mechanical Turk for Computer Vision

Course

*New*: The slides for the course are attached below.

The short course at CVPR 2010 will teach how to use Mechanical Turk for a computer vision project.

When: Friday, June 18, 2010, 8:30 am - 12:30 pm.
Presented by: Alexander Sorokin, Fei-Fei Li

The course will follow the cover these topics:

1. Introduction to crowdsourcing
2. Tools for crowdsourcing
3. Hand-on examples
4. Issues in crowdsourcing:
   - How to define a task
   - How to manage work
   - Quality control techniques
   - How much to pay
5. Case studies
6. Discussion and open issues

Attachments (2)

example1.ppt - on Jun 29, 2010 10:51 AM by Alexander Sorokin (version 1)

Download
SA1: How to use Mechanical Turk for Behavioral Research

Winter Mason (winteram@yahoo-inc.com) and Siddharth Suri (suri@yahoo-inc.com)
9:00 AM - 12:00 PM

In this tutorial, we will demonstrate how to conduct behavioral research on Amazon’s Mechanical Turk. We will begin by discussing the four main advantages to using Mechanical Turk as a platform for running online studies: access to a large pool of participants, diversity of the participants, low cost of running studies, and faster research cycle. We will outline the fundamental components of a job on Mechanical Turk and discuss the features of the marketplace, including who is doing the work. We will describe how to run three kinds of studies on Mechanical Turk: surveys, experiments with random assignment, and synchronous experiments. We will demonstrate the mechanics of putting a task on Mechanical Turk by creating a survey, posting the job to Mechanical Turk, reviewing the responses and paying the workers. Finally, we will discuss methods for quality assurance and ethical issues surrounding Mechanical Turk.

Winter Mason received a B.S. in Psychology from University of Pittsburgh in 1999 and a Ph.D. in Cognitive Science and Social Psychology from Indiana University in 2007. Since then he has worked at Yahoo! Research in the Human Social Dynamics group.

Siddharth Suri joined the Human & Social Dynamics group at Yahoo! Research in August 2008. Prior to that he was a postdoctoral associate in the computer science department at Cornell University. He earned his Ph.D. in computer and information science from the University of Pennsylvania in January 2007.
Creating Speech and Language Data With Amazon’s Mechanical Turk

Note: We welcome papers on ALL crowd source tools (Amazon Mechanical Turk, CrowdFlower, Herd It, etc.)

Amazon's Mechanical Turk is an online marketplace for work that allows you to pay people small sums of money to do "Human Intelligence Tasks" or HITs. Tasks include anything from labeling images, to listening to short pieces of audio, to researching topics on the internet, to scrubbing database records.

A number of recent papers have evaluated the effectiveness of using Mechanical Turk to create annotated data for natural language processing applications. Mechanical Turk's low cost, scalable workforce opens new possibilities for annotating speech and text, and has the potential to dramatically change how we create data for human language technologies. Open questions include:

- How can we ensure high quality annotations?
- What tools are available for obtaining complex annotations?
- What types of annotations and evaluations are possible when the cost is dramatically reduced?

This work will explore uses of Mechanical Turk in several ways:

- **Shared task**: What can you do with $100 and Mechanical Turk? Participants will be given a budget to spend on Mechanical Turk and submit papers describing the results of their experience.
- **General papers**: These papers will explore general issues with using Mechanical Turk for language processing research.

Discussion Group
Running experiments on Amazon Mechanical Turk

Gabriele Paolacci*
Advanced School of Economics, Ca’ Foscari University of Venice

Jesse Chandler
Woodrow Wilson School of Public and International Affairs, Princeton University

Panagiotis G. Ipeirotis
Leonard N. Stern School of Business, New York University

Abstract

Although Mechanical Turk has recently become popular among social scientists as a source of experimental data, doubts may linger about the quality of data provided by subjects recruited from online labor markets. We address these potential concerns by presenting new demographic data about the Mechanical Turk subject population, reviewing the strengths of Mechanical Turk relative to other online and offline methods of recruiting subjects, and comparing the magnitude of effects obtained using Mechanical Turk and traditional subject pools. We further discuss some additional benefits such as the possibility of longitudinal, cross cultural and prescreening designs, and offer some advice on how to best manage a common subject pool.

Keywords: experimentation, online research

1 Introduction

Mechanical Turk started in 2005 as a service to “crowd-democratize” online surveys, at least as representative of the U.S. population as traditional subject pools. Further, we show that it is shift-
Conducting Behavioral Research on Amazon’s Mechanical Turk

Winter Mason and Siddharth Suri
Yahoo! Research

Amazon’s Mechanical Turk is an online labor market where requesters post jobs and workers choose which jobs to do for pay. The central purpose of this paper is to demonstrate how to use this website for conducting behavioral research and lower the barrier to entry for researchers who could benefit from this platform. We describe general techniques that apply to a variety of types of research and experiments across disciplines. We begin by discussing some of the advantages of doing experiments on Mechanical Turk, such as easy access to a large, stable, and diverse subject pool, the low cost of doing experiments and faster iteration between developing theory and executing experiments. We will discuss how the behavior of workers compares to experts and to laboratory subjects. Then, we illustrate the mechanics of putting a task on Mechanical Turk including recruiting subjects, executing the task, and reviewing the work that was submitted. We also provide solutions to common problems that a researcher might face when executing their research on this platform including techniques for conducting synchronous experiments, methods to ensure high quality work, how to keep data private, and how to maintain code security.
The Online Laboratory:
Conducting Experiments in a Real Labor Market*

John J. Horton†  David G. Rand‡  Richard J. Zeckhauser§

Abstract

Online labor markets have great potential as platforms for conducting experiments, as they provide immediate access to a large and diverse subject pool and allow researchers to conduct randomized controlled trials. We argue that online experiments can be just as valid—both internally and externally—as laboratory and field experiments, while requiring far less money and time to design and to conduct. In this paper, we first describe the benefits of conducting experiments in online labor markets; we then use one such market to replicate three classic experiments and confirm their results. We confirm that subjects (1) reverse decisions in response to how a decision-problem is framed, (2) have pro-social preferences (value payoffs to others positively), and (3) respond to priming by altering their choices. We also conduct a labor supply field experiment in which we confirm that workers have upward sloping labor supply curves. In addition to reporting these results, we discuss the unique threats to validity
Types of Collective Intelligence

• **Crowd Mining**: Achieve desired outcomes by mining the observed behaviors of people
  – Amazon recommendations
  – Spam detection
  – Google
Explore flu trends around the world

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

Download world flu activity data  -  Animated flu trends for Google Earth  -  Compare flu trends across regions in Public Data Explorer
United States Flu Activity

Influenza estimate

- Blue line: Google Flu Trends estimate
- Orange line: United States data

United States: Influenza-like illness (ILI) data provided publicly by the U.S. Centers for Disease Control.
Types of Collective Intelligence

• **Crowd Mining**: Achieve desired outcomes by mining the observed behaviors of people
  – Amazon recommendations
  – Spam Detection
  – Google
  – Google Flu Trends
Types of Collective Intelligence

• **Crowd Mining**: Achieve desired outcomes by mining the observed behaviors of people
  – Amazon recommendations
  – Spam Detection
  – Google
  – Google Flu Trends
  – Google Translate
Showing results for machine learning.
Search instead for machine learning.

Machine learning - Wikipedia, the free encyclopedia
Machine learning, a branch of artificial intelligence, is a scientific discipline concerned with the design and development of algorithms that allow
en.wikipedia.org/wiki/Machine_learning - Cached - Similar

List of machine learning algorithms
Machine Learning (journal)
Category:Machine learning

More results from wikipedia.org »

Introduction to Machine Learning
Jun 19, 2010 – From this page you can download a draft of notes I used for a Stanford course on Machine Learning. Although I have tried to eliminate errors ...
robotics.stanford.edu/~nilsson/mlbook.html - Cached
Types of Collective Intelligence

• **Crowd Mining**: Achieve desired outcomes by mining the observed behaviors of people
  – Amazon recommendations
  – Spam Detection
  – Google
  – Google Flu Trends
  – Google Translate
  – Spelling correction
Perspectives on Collective Intelligence: Framework Papers
ABSTRACT

People with shared interests are using the Internet to solve problems, accomplish tasks, and create resources that would be well beyond the reach of any one person or organization. The Internet is being used to create virtual libraries, factor large numbers, organize massive volunteer efforts, and filter information in a collaborative fashion. The ability to leverage the efforts of large numbers of networked users has important economic, social, and political consequences. This phenomenon is important to policy makers because it can potentially be used to leverage scarce taxpayer dollars and promote applications of the information infrastructure.

Leveraging Cyberspace

Thomas A. Kalil, National Economic Council

"Give me a lever long enough and a place to stand, and I will move the Earth."

— Archimedes

The rapid growth in the ubiquity and functionality of the Internet is amazing. The Internet now connects 10 million computers, tens of millions of users, and more than 100 countries. At its current rate of growth, the Internet will connect 100 million computers by the year 2000. Because anyone with a computer and a connection to the Internet can publish, the global information space is also growing rapidly. Developers of search engines such as Altavista and Lycos believe that the Web currently contains 30–50 million pages of information, or 200 to 330 Gbytes of text. At current growth rates, the Web could surpass the 29 Tbytes of the Library of Congress in two years [1].

In addition to allowing anyone to publish, the open architecture of the Internet also allows anyone to add to its func.
Perspectives on Collective Intelligence: Leveraging Cyberspace – Kalil 1996

- Can dependencies between parts of the task be eliminated or managed?
- What will motivate people to participate?
- Is there part of the task that must be centrally administered?
- Does the initiative demonstrate increasing returns?
Perspectives on Collective Intelligence: Wisdom of Crowds – Surowiecki 2004

• Requires participant:
  – Diversity
  – Independence
  – Decentralization
  – Aggregation
Perspectives on Collective Intelligence: Malone, Laubacher and Dellarocas

- Who is doing it?
- Why are they doing it?
- How is it being done?
- What is being done?
Perspectives on Collective Intelligence: Malone, Laubacher and Dellarocas

• Who is doing it?
  – Organization / Authority selects
  – Crowd: Individuals select

• Why are they doing it?
  – Money
  – Love (e.g., enjoyment, socializing)
  – Glory

• What is being done?
  – Create
  – Decide
Perspectives on Collective Intelligence: Malone, Laubacher and Dellarocas

• How is it being done?
  – Create
    • Collection (incl. Contest)
    • Collaboration
  – Decide
    • Group
      – Voting, Consensus, Averaging, Markets
    • Individual
      – Markets
      – Social Network

3 July 2011
Copyright 2011 Haym Hirsh
Perspectives on Collective Intelligence: Doan, Ramakrishnan, and Halevy

• How to recruit and retain users?
• What contributions can users make?
• How to combine user contributions to solve the target problem?
• How to evaluate users and their contributions?
Perspectives on Collective Intelligence: Quinn and Bederson

• Motivation:
  – Pay (AMT)
  – Altruism (search for Jim Gray)
  – Fun (ESP Game)
  – Reputation (Threadless)
  – Implicit Work [Coercion] (ReCAPTCHA)
  – [Economical models (e.g., game theory)]
The Influence Limiter: Provably Manipulation-Resistant Recommender Systems

Paul Resnick
University of Michigan
School of Information
presnick@umich.edu

Rahul Sami
University of Michigan
School of Information
rsami@umich.edu

ABSTRACT
An attacker can draw attention to items that don’t deserve that attention by manipulating recommender systems. We describe an influence-limiting algorithm that can turn existing recommender systems into manipulation-resistant systems. Honest reporting is the optimal strategy for raters who wish to maximize their influence. If an attacker can create only a bounded number of shills, the attacker can mislead only a small amount. However, the system eventually makes full use of information from honest, informative raters. We describe both the influence limits and the information loss incurred due to those limits in terms of information-theoretic concepts of loss functions and entropies.

Categories and Subject Descriptors
I.2.6 [Computing Methodologies]: Artificial Intelligence—Learning

General Terms
Algorithms, Reliability

Keywords
Recommender systems, manipulation-resistance, shilling

1. INTRODUCTION
Content posted on the Internet is not of uniform quality, nor is it equally interesting to different audiences. Recommender systems guide people to items they are likely to like, based on their own and other people’s subjective re-

into a multi-billion dollar advertising marketplace. But to the extent that people rely on recommender systems of various kinds to guide their attention, there are also natural incentives for promoters to manipulate the recommendations. An attacker may rate strategically rather than honestly and may introduce multiple entities, sometimes called sybils, to rate on behalf of the attacker.

We offer a manipulation-resistance algorithm, called the Influence Limiter, that can be overlaid on existing recommender algorithms. Consider the predictions about whether a particular target person will like various items. Each rater begins with a very low but non-zero reputation score. The current reputation limits the influence she can have on the prediction for the next item. Eventually, the target person indicates whether he likes the item and the raters who contributed to predicting whether the target would like it gain or lose reputation. The more that a rating implies a change in the prediction for the target, the greater the potential change in the rater’s reputation score. A rater who simply goes along with the previous change will have no impact and thus get no change in her reputation.

The Influence Limiter has several desirable properties. First, in order to maximize the expected reputation score of a single rater endowed with some information about the target’s likely response to the items, the optimal strategy is to induce predictions that accurately reveal that rater’s information about the items. If the underlying recommendation algorithm is making optimal use of ratings, this implies that entering honest ratings is optimal. An important special case is that a rater who has not interacted with an item, and therefore has no information about the target’s likely response to it, can only lose reputation in expectation by giving a rating.
Perspectives on Collective Intelligence: Quinn and Bederson

• Reliability
  – Output agreement (ESP Game)
  – Input agreement (Tag-a-tune)
  – Economical models (e.g., game theory)
  – Defensive task design (AMT)
  – Reputation system (AMT, Wikipedia)
  – Redundancy/voting (AMT)
  – Ground truth seeding (AMT)
  – Statistical filtering/aggregating
  – Multilevel review (AMT)
  – Expert review
  – Automatic checking (fold.it)
Perspectives on Collective Intelligence: Quinn and Bederson

• Aggregation
  – Collection (Wikipedia, reCaptcha)
  – Statistical processing (e.g., averaging) (IEM, Wisdom of the Crowds, Kasparov against the World)
  – Iterative improvement (TurkIt, Soylent)
  – Active learning
  – Search
  – [Genetic algorithm]
  – None
TurKit: Human Computation Algorithms on Mechanical Turk

Greg Little\textsuperscript{1}, Lydia B. Chilton\textsuperscript{2}, Max Goldman\textsuperscript{1}, Robert C. Miller\textsuperscript{1}

\textsuperscript{1}MIT CSAIL
\{glittle, maxg, rcm\}@mit.edu

\textsuperscript{2}University of Washington
hmslydia@cs.washington.edu

ABSTRACT

Mechanical Turk provides an on-demand source of human computation. This provides a tremendous opportunity to explore algorithms which incorporate human computation as a function call. However, various systems challenges make this difficult in practice, and most uses of Mechanical Turk post large numbers of independent tasks. TurKit is a toolkit for prototyping and exploring truly algorithmic human computation, while maintaining a straightforward imperative programming style. We present the crash-and-rerun programming model that makes TurKit possible, along with a variety of applications for human computation algorithms. We also present a couple case studies of TurKit used for real experiments outside our lab.

ACM Classification: H5.2 [Information interfaces and presentation]: User Interfaces. - Prototyping.

General terms: Algorithms, Design, Experimentation

Keywords: Human computation, Mechanical Turk, toolkit

INTRODUCTION

Amazon’s Mechanical Turk (MTurk) is a popular web ser-

general, this paper considers human computation algorithms, where an algorithm coordinates the contributions of humans toward some goal.

Figure 1: Naturally, a programmer wants to write an algorithm to help them visit New York City. TurKit lets them use Mechanical Turk as a function call to generate ideas and compare them.

```javascript
ideas = []
for (var i = 0; i < 5; i++) {
  idea = mturk.prompt(
    "What's fun to see in New York City?"
    Ideas so far: " + ideas.join(", ")
  )
  ideas.push(idea)
}

ideas.sort(function (a, b) {
  v = mturk.vote("Which is better?", [a, b])
  return v == a ? -1 : 1
})
```
“You (misspelled) (several) (words). Please spellcheck your work next time. I also notice a few grammatical mistakes. Overall your writing style is a bit too phoney. You do make some good (points), but they got lost amidst the (writing). (signature)

Highlighted words should be: “flowery”, “get”, “verbiage”, and “B-.”
Soylent: A Word Processor with a Crowd Inside

Michael S. Bernstein\textsuperscript{1}, Greg Little\textsuperscript{1}, Robert C. Miller\textsuperscript{1}, Björn Hartmann\textsuperscript{2}, Mark S. Ackerman\textsuperscript{3}, David R. Karger\textsuperscript{1}, David Crowell\textsuperscript{1}, Katrina Panovich\textsuperscript{1}

\textsuperscript{1} MIT CSAIL
Cambridge, MA
\{msbernst, glittle, rcm, karger, dcrowell, kp\}@csail.mit.edu

\textsuperscript{2} Computer Science Division
University of California, Berkeley
Berkeley, CA
bjoern@cs.berkeley.edu

\textsuperscript{3} Computer Science & Engineering
University of Michigan
Ann Arbor, MI
ackerm@umich.edu

ABSTRACT
This paper introduces architectural and interaction patterns for integrating crowdsourced human contributions directly into user interfaces. We focus on writing and editing, complex endeavors that span many levels of conceptual and pragmatic activity. Authoring tools offer help with pragmatics, but for higher-level help, writers commonly turn to other people. We thus present Soylent, a word processing interface that enables writers to call on Mechanical Turk workers to shorten, proofread, and otherwise edit parts of their documents on demand. To improve worker quality, we introduce the Find-Fix-Verify crowd programming pattern, which splits tasks into a series of generation and review stages. Evaluation studies demonstrate the feasibility of crowdsourced editing and investigate questions of reliability, cost, wait time, and work time for edits.

ACM Classification: H5.2 [Information interfaces and presentation]: User Interfaces. - Graphical user interfaces.
General terms: Design, Human Factors
Keywords: Outsourcing, Mechanical Turk, Crowdsourcing

edits on Wikipedia [13]. Writing is no exception [7]: we commonly recruit friends and colleagues to help us shape and polish our writing. But we cannot always rely on them: colleagues do not want to proofread every sentence we write, cut a few lines from every paragraph in a ten-page paper, or help us format thirty ACM-style references.

As a step toward integrating this human expertise permanently into our writing tools, we present Soylent, a word processing interface that utilizes crowd contributions to aid complex writing tasks ranging from error prevention and paragraph shortening to automation of tasks like citation searches and tense changes. We hypothesize that crowd workers with a basic knowledge of written English can support both novice and expert writers. These workers perform tasks that the writer might not, such as scrupulously scanning for text to cut, or updating a list of addresses to include a zip code. They can also solve problems artificial intelligence cannot, for example flat errors that the word processor does not catch.

Soylent aids the writing process by integrating paid crowd
Perspectives on Collective Intelligence: Quinn and Bederson

• Human Skill
• Process Order
• Task Request Cardinality
Looking Ahead:
Multi-Layered Collective Intelligence
Customer Reviews
The Wisdom of Crowds

199 Reviews
Average Customer Review
5 star: (86)
4 star: (53)
3 star: (27)
2 star: (20)
1 star: (13)

Share your thoughts with other customers
Create your own review

Search Customer Reviews

Search for: 

Only search this product's reviews

The most helpful favorable review
251 of 276 people found the following review helpful:

⭐⭐⭐⭐⭐ A Counter-Intuitive Notion
In 1906, Francis Galton, known for his work on statistics and heredity, came across a weight-judging contest at the West of England Fat Stock and Poultry Exhibition. This encounter was to challenge the foundations of his life's study.
TRANSCRIPTION MADE FAST AND EASY.

CASTINGWORDS TRANSCRIPTION SERVICES:

QUALITY IS KING
We catch the difficult words, do the research, and produce rocking transcripts.

CLEAR PRICING
We price per minute of audio - so you know how much you'll be paying now. No surprises.

SIMPLE TRANSCRIPT DELIVERY
Transcripts are emailed to you - and can be picked up here at the site.

LOTS OF FORMATS
Transcripts are delivered in plain text, HTML, and MS Word formats.

HUMAN TRANSCRIPTION
All transcription is done by people, not machines. All transcripts are checked many times, by many people.

LIVE ORDER TRACKING
You always know what is happening with your transcription order.

We were constantly surprised that phrases or names that other companies missed, from "tabula rasa" to "turducken" (it's a Cajun specialty), were identified and spelled correctly.

WALL STREET JOURNAL

NEWS & UPDATES!
NOVEMBER 4TH 2010: You can now upgrade your order (and add timestamps and difficult audio) all by yourself. Just login to your account and click "Upgrade"
Looking Ahead:
Programming the Social Computer
TurKit: Human Computation Algorithms on Mechanical Turk

Greg Little\textsuperscript{1}, Lydia B. Chilton\textsuperscript{2}, Max Goldman\textsuperscript{1}, Robert C. Miller\textsuperscript{1}

\textsuperscript{1}MIT CSAIL
\{glittle, maxg, rcm\}@mit.edu

\textsuperscript{2}University of Washington
hmslydia@cs.washington.edu

ABSTRACT
Mechanical Turk provides an on-demand source of human computation. This provides a tremendous opportunity to explore algorithms which incorporate human computation as a function call. However, various systems challenges make this difficult in practice, and most uses of Mechanical Turk post large numbers of independent tasks. TurKit is a toolkit for prototyping and exploring truly algorithmic human computation, while maintaining a straightforward imperative programming style. We present the crash-and-rerun programming model that makes TurKit possible, along with a variety of applications for human computation algorithms. We also present a couple case studies of TurKit used for real experiments outside our lab.

ACM Classification: H5.2 [Information interfaces and presentation]: User Interfaces. - Prototyping.

General terms: Algorithms, Design, Experimentation

Keywords: Human computation, Mechanical Turk, toolkit

INTRODUCTION
Amazon’s Mechanical Turk (MTurk) is a popular web service.
CrowdForge: Crowdsourcing Complex Work

Aniket Kittur, Boris Smus, Robert E. Kraut

February 1, 2011
CMU-HCII-11-100

Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3890

Abstract: Micro-task markets such as Amazon’s Mechanical Turk represent a new paradigm for accomplishing work, in which employers can tap into a large population of workers around the globe to accomplish tasks in a fraction of the time and money of more traditional methods. However, such markets typically support only simple, independent tasks, such as labeling an image or judging the relevance of a search result. Here we present a general purpose framework for accomplishing complex tasks using micro-task markets. Our approach is inspired by the MapReduce framework for distributed processing and provides a scaffolding for complex human computation tasks. We describe our general framework, a web-based prototype, and case studies on article writing and decision making that demonstrate the benefits of the approach.
CrowdFlow: Integrating Machine Learning with Mechanical Turk for Speed-Cost-Quality Flexibility

Alexander J. Quinn\textsuperscript{1}, Benjamin B. Bederson\textsuperscript{1,2,3}, Tom Yeh\textsuperscript{3}, Jimmy Lin\textsuperscript{1,2}

University of Maryland
Human-Computer Interaction Lab
Department of Computer Science\textsuperscript{1} || iSchool\textsuperscript{2} || Institute for Advanced Computer Studies\textsuperscript{3}
College Park, MD 20742 USA

aq@cs.umd.edu, bederson@cs.umd.edu, tomyeh@umiacs.umd.edu, jimmylin@umd.edu

Abstract
Humans and machines have competing strengths for tasks such as natural language processing and image understanding. Whereas humans do these things naturally with potentially high accuracy, machines offer greater speed and flexibility. CrowdFlow is our toolkit for a model for blending the two in order to attain tighter control over the inherent tradeoffs in speed, cost and quality. With CrowdFlow, humans and machines work together to do a set of tasks at a user-specified point in the tradeoff space. They work symbiotically, with the humans providing training data to the machine while the machine provides first cut results to the humans to save effort in cases where the machine’s answer was already correct. The CrowdFlow toolkit can be considered as a generalization of our other domain-specific efforts aimed at enabling cloud computing services using a variety of computational resources to achieve various tradeoff points.

1. Introduction
There is a large set of problems that can be solved either by human computation or machine learning. These include recognizing the faces of missing children in surveillance videos, translating documents between languages, or summarizing the participants in collaboration with machine translation systems to translate books - resulting in translations which are between pure machine and pure expert (bilingual) human solutions [2][9]. Although MonoTrans is not a GWAP, it is similar in that it yields relatively high quality and low cost, at the expense of time. Naive uses of AMT provide faster turnaround time (and a greater range of potential applications) at the expense of cost and quality.

The principal goal of this research is to discover strategies that combine the strengths of humans and machines, in order to move past these two rigid extremes. We would like to give developers the flexibility to choose any point in the speed-cost-quality spectrum, and automatically combine human and machine resources to achieve the desired balance. To do so in a general way is an ambitious goal.

We are developing a framework for blending capabilities of humans and machines in order to enable much greater flexibility with respect to speed-cost-quality. The strategy blends the flexibility of Amazon Mechanical Turk with machine learning (or other automated methods).

To demonstrate the framework, we built the CrowdFlow toolkit, a programming library for distributing tasks between AMT and
CrowdDB: Answering Queries with Crowdsourcing

Michael J. Franklin
AMPLab, UC Berkeley
franklin@cs.berkeley.edu

Donald Kossmann
Systems Group, ETH Zurich
donaldk@inf.ethz.ch

Tim Kraska
AMPLab, UC Berkeley
kraska@cs.berkeley.edu

Sukriti Ramesh
Systems Group, ETH Zurich
ramess@student.ethz.ch

Reynold Xin
AMPLab, UC Berkeley
rxin@cs.berkeley.edu

ABSTRACT

Some queries cannot be answered by machines only. Processing such queries requires human input for providing information that is missing from the database, for performing computationally difficult functions, and for matching, ranking, or aggregating results based on fuzzy criteria. CrowdDB uses human input via crowdsourcing to process queries that neither database systems nor search engines can adequately answer. It uses SQL both as a language for posing complex queries and as a way to model data. While CrowdDB leverages many aspects of traditional database systems, there are also important differences. Conceptually, a major change is that the traditional closed-world assumption for query processing does not hold for human input. From an implementation perspective, human-oriented query operators are needed to solicit, integrate, and cleanse crowdsourced data. Furthermore, performance and cost depend on a number of new factors including worker affinity, training, fatigue, motivation and location. We describe the design of CrowdDB, report on an initial set of experiments using Amazon Mechanical Turk, and outline important avenues for future work in the development of crowdsourced query processing systems.

assumptions about the correctness, completeness and unambiguity of the data they store. When these assumptions fail to hold, relational systems will return incorrect or incomplete answers to user questions, if they return any answers at all.

1.1 Power to the People

One obvious situation where existing systems produce wrong answers is when they are missing information required for answering the question. For example, the query:

```
SELECT market_capitalization FROM company
WHERE name = "I.B.M."
```

will return an empty answer if the company table instance in the database at that time does not contain a record for “I.B.M.”. Of course, in reality, there are many reasons why such a record may be missing. For example, a data entry mistake may have omitted the I.B.M. record or the record may have been inadvertently deleted. Another possibility is that the record was entered incorrectly, say, as “I.B.N.”

Traditional systems can erroneously return an empty result even
Crowdsourced Databases: Query Processing with People

Adam Marcus, Eugene Wu, David R. Karger, Samuel Madden, Robert C. Miller
MIT CSAIL
{marcua, sirrice, karger, madden, rcm}@csail.mit.edu

ABSTRACT

Amazon’s Mechanical Turk (“MTurk”) service allows users to post short tasks (“HITs”) that other users can receive a small amount of money for completing. Common tasks on the system include labelling a collection of images, combining two sets of images to identify people which appear in both, or extracting sentiment from a corpus of text snippets. Designing a workflow of various kinds of HITs for filtering, aggregating, sorting, and joining data sources together is common, and comes with a set of challenges in optimizing the cost per HIT, the overall time to task completion, and the accuracy of MTurk results. We propose Qurk, a novel query system for managing these workflows, allowing crowd-powered processing of relational databases. We describe a number of query execution and optimization challenges, and discuss some potential solutions.

1. INTRODUCTION

Amazon’s Mechanical Turk service (https://www.mturk) processing. For example, given a database storing a table of images, a user might want to query for images of flowers, generating a HIT per image to have turkers perform the filter task. Several challenges arise in processing such a workflow. First, each HIT can take minutes to return, requiring an asynchronous query executor. Second, in addition to considering time, a crowdworker-aware optimizer must consider monetary cost and result accuracy. Finally, the selectivity of operators can not be predicted a priori, requiring an adaptive approach to query processing.

In this paper, we propose Qurk, a crowdworker-aware database system which addresses these challenges. Qurk can issue HITs that extract, order, filter, and join complex datatypes, such as images and large text blobs. While we describe Qurk here using the language of MTurk (turkers and HITs), and our initial prototype runs on MTurk, we aim for Qurk to be crowd-platform-independent. Future versions of Qurk will compile tasks for different kinds of crowds with different interfaces and incentive systems. Qurk is a new system in active development; we describe our vision for the
Looking Ahead:
Understanding People
Breaking Monotony with Meaning: Motivation in Crowdsourcing Markets

Dana Chandler\textsuperscript{a,\ast}, Adam Kapelner\textsuperscript{b,\ast\ast}

\textsuperscript{a}University of Chicago
Becker Center on Chicago Price Theory
5807 South Woodlawn, STE 370
Chicago, IL 60637

\textsuperscript{b}The Wharton School of the University of Pennsylvania,
3730 Walnut Street
Philadelphia, PA 19104

First Draft: 25 Jan 2010
This Draft: 2 May 2010

Abstract
We conduct a natural field experiment that explores the relationship between the “meaningfulness” of a task and people’s willingness to work. Our study uses workers from Amazon’s Mechanical Turk (MTurk), an online marketplace for task-based work. All participants are given an identical task of labeling medical images. However, the task is presented differently depending on treatment. Subjects assigned to the meaningful treatment are told they would be helping researchers label tumor cells, whereas subjects in the zero-context treatment are not told the purpose of their task and only told that they would be labeling “objects of interest”. Our experimental design specifically hires US and Indian workers in order to test for heterogeneous effects. We find that US, but not
Scientists constantly producing images that need to be analyzed

- Scientists are always producing images
- It takes lots of time to analyze the images
- People can be trained to do it

We’ll train you to identify tumor cells in images
More people were induced to work for a meaningful task

Proportion induced to work

US and India

United States

Treatment effect by Country

India

Zero-context

Meaningful
Financial Incentives and the “Performance of Crowds”

Winter Mason
Yahoo! Research
111 West 40th Street
New York, NY 10018
+1 (212) 571-8124
winteram@yahoo-inc.com

Duncan J. Watts
Yahoo! Research
111 West 40th Street
New York, NY 10018
+1 (212) 571-8126
djw@yahoo-inc.com

ABSTRACT

The relationship between financial incentives and performance, long of interest to social scientists, has gained new relevance with the advent of web-based “crowd-sourcing” models of production. Here we investigate the effect of compensation on performance in the context of two experiments, conducted on Amazon’s Mechanical Turk (AMT). We find that increased financial incentives increase the quantity, but not the quality, of work performed by participants, where the difference appears to be due to an “anchoring” effect: workers who were paid more also perceived the value of their work to be greater, and thus were no more motivated than workers paid less. In contrast with compensation levels, we find the details of the compensation scheme do matter—specifically, a “quota” system results in better work for less pay than an equivalent “piece rate” system. Although counterintuitive, these findings are consistent with previous laboratory studies, and may have real-world analogs as well.

Flickr and Del.icio.us. One important sub-class of peer production is a phenomenon known as “crowd-sourcing” [4, 5] in which potentially large jobs are broken into many small tasks that are then outsourced directly to individual workers via public solicitation. Workers sometimes work for free, motivated either out of intrinsic enjoyment [3] or some form of social reward [6]; however, successful examples of volunteer crowd sourcing have proven difficult to replicate, in part because arbitrary tasks tend not to be intrinsically enjoyable, and in part because social rewards are often highly context specific. As a result, crowd sourcing increasingly uses financial compensation, often in the form of micro-payments on the order of a few cents per task. This model has the advantage that it is more replicable than models based on intrinsic or social rewards; yet it can still accomplish tasks quickly and cheaply. As a result, paid crowd-sourcing has elicited considerable interest as an alternative mode of production to traditional firms [4]. Nevertheless, the success of any given enterprise still depends to some extent on the ability of would-be “employers” to attract the appropriate workers and motivate them.
You’re Hired! An Examination of Crowdsourcing Incentive Models in Human Resource Tasks

Christopher G. Harris
Informatics Program
The University of Iowa
Iowa City, IA 52242

christopher-harris@uiowa.edu

ABSTRACT

Many human resource tasks, such as screening a large number of job candidates, are labor-intensive and rely on subjective evaluation, making them excellent candidates for crowdsourcing. We conduct several experiments using the Amazon Mechanical Turk platform to conduct resume reviews. We then apply several incentive-based models and examine their effects. Next, we assess the accuracy measures of our incentive models against a gold standard and ascertain which incentives provide the best results. We find that some incentives actually encourage quality if the task is designed appropriately.

Categories and Subject Descriptors
H.3.4 [Information Storage and Retrieval]: Systems and Software - Performance Evaluation

General Terms
Measurement, Design, Experimentation, Human Factors

Keywords
... 

Frequently the task of hiring mid-level employees and above is outsourced to executive search firms. These outside recruiters typically charge around one third of the annual base salary of a newly-hired employee. Therefore an inexpensive method of examining resumes can benefit employers or outside search firms cut costs substantially if this activity can be done effectively.

Crowdsourcing tools such as Amazon’s Mechanical Turk (AMT) show considerable promise in having simple yet tedious tasks executed rapidly. These platforms provide a legion of available Internet workers to complete HITs (Human Intelligence Tasks) in exchange for micro-payments – precisely the type of activity that can help HR recruiters narrow a pile of resumes to only those of interest. By dividing a tedious task among a large number of participants, a company can quickly and inexpensively execute tasks in a short timeframe, often within 24 hours.

Our objective is to examine how platforms such as AMT can do well with the types of subjective evaluations computers cannot perform well. Additionally, we wish to examine the role incentives play in aligning the worker’s needs with those of the requester in this anonymous environment.
Research Article

God Is Watching You

Priming God Concepts Increases Prosocial Behavior in an Anonymous Economic Game

Azim F. Shariff and Ara Norenzayan

University of British Columbia, Vancouver, British Columbia, Canada

ABSTRACT—We present two studies aimed at resolving experimentally whether religion increases prosocial behavior in the anonymous dictator game. Subjects allocated more money to anonymous strangers when God concepts were implicitly activated than when neutral or no concepts were activated. This effect was at least as large as that obtained when concepts associated with secular moral institutions were primed. A trait measure of self-reported religiosity did not seem to be associated with prosocial behavior. We discuss different possible mechanisms that may underlie this effect, focusing on the hypotheses that the religious prime had an ideomotor effect on generosity or that it activated a felt presence of supernatural watchers. We then discuss implications for theories positing religion as a facilitator of the emergence of early large-scale societies of cooperators.

Sosis and Ruffle (2004) examined levels of generosity in an experimental cooperative pool game in religious and secular kibbutzim in Israel and found higher levels of cooperation in the religious ones, and the highest levels among religious men who engaged in daily communal prayer. Batson and his colleagues (Batson et al., 1989; Batson, Schoenrade, & Ventis, 1993) have shown that although religious people report more explicit willingness to care for others than do nonreligious people, controlled laboratory measures of altruistic behavior often fail to corroborate this difference. Furthermore, when studies demonstrate that helpfulness is higher among more devoted people, this finding is typically better explained by egoistic motives such as seeking praise or avoiding guilt, rather than by higher levels of compassion or by a stronger motivation to benefit other people.

However insightful these findings are, research on religion and prosocial behavior has been limited by its overwhelming
Looking Ahead:
Seeking Reliability
Rethinking the ESP Game

Ingmar Weber
EPFL
Lausanne, Switzerland
ingmar.weber@epfl.ch

Stephen Robertson
Microsoft Research
Cambridge, UK
ser@microsoft.com

Milan Vojnović
Microsoft Research
Cambridge, UK
milanv@microsoft.com

September 2008

Technical Report
MSR-TR-2008-132

Abstract — The ESP Game [15] was designed to harvest human intelligence to assign labels to images - a task which is still difficult for even the most advanced systems in image processing [2, 8]. However, the ESP Game as it is currently implemented encourages players to assign “obvious” labels, which are most likely to lead to an agreement with the partner. But these labels can often be deduced from the labels already present using an appropriate language model and such labels therefore add only little information to the system.

We present a language model which, given enough instances of labeled images as training data, can assign probabilities to the next label to be added. This model is then used in a program, which plays the ESP game without looking at the image. Even without any understanding of the actual image, the program manages to agree with the randomly assigned human partner on a label for 69% of all images, and for 81% of images which have at least one “off-limits” term assigned to them.

We then show how, given any generative probabilistic model, the scoring system for the ESP game can be redesigned to encourage users to add less predictable labels, thereby leading to a collection of informative, high entropy tag sets. Finally, we discuss a number of other possible redesign options to improve the quality of the collected labels.
crowdsourcing reliability

Crowdsourcing quality

Crowdsourcing reputation

Graph showing trends from 2006 to 2010 for crowdsourcing reliability, quality, and reputation.
Adversarial Classification

Nilesh Dalvi  Pedro Domingos  Mausam  Sumit Sanghavi  Deepak Verma
Department of Computer Science and Engineering
University of Washington, Seattle
Seattle, WA 98195-2350, U.S.A.
{nilesh,pedrod,mausam,sanghavi,deepak}@cs.washington.edu

ABSTRACT
Essentially all data mining algorithms assume that the data-generating process is independent of the data miner’s activities. However, in many domains, including spam detection, intrusion detection, fraud detection, surveillance and counter-terrorism, this is far from the case: the data is actively manipulated by an adversary seeking to make the classifier produce false negatives. In these domains, the performance of a classifier can degrade rapidly after it is deployed, as the adversary learns to defeat it. Currently, the only solution to this is repeated, manual, ad hoc reconstruction of the classifier. In this paper we develop a formal framework and algorithms for this problem. We view classification as a game between the classifier and the adversary, and produce a classifier that is optimal given the adversary’s optimal strategy. Experiments in a spam detection domain show that this approach can greatly outperform a classifier learned in the standard way, and (within the parameters of the problem) automatically adapts the classifier to the adversary’s evolving manipulations.

Categories and Subject Descriptors
H.2.8 [Database Management]: Database Applications—data mining; I.2.6 [Artificial Intelligence]: Learning—concept learning, induction, computer learning; J.5.1 [Dat...
The Influence Limiter: Provably Manipulation-Resistant Recommender Systems

Paul Resnick
University of Michigan
School of Information
presnick@umich.edu

Rahul Sami
University of Michigan
School of Information
rsami@umich.edu

ABSTRACT
An attacker can draw attention to items that don't deserve that attention by manipulating recommender systems. We describe an influence-limiting algorithm that can turn existing recommender systems into manipulation-resistant systems. Honest reporting is the optimal strategy for raters who wish to maximize their influence. If an attacker can create only a bounded number of shills, the attacker can mislead only a small amount. However, the system eventually makes full use of information from honest, informative raters. We describe both the influence limits and the information loss incurred due to these limits in terms of information-theoretic concepts of loss functions and entropies.

Categories and Subject Descriptors
I.2.6 [Computing Methodologies]: Artificial Intelligence—Learning

General Terms
Algorithms, Reliability

Keywords
Recommender systems, manipulation-resistance, shilling

1. INTRODUCTION
Content posted on the Internet is not of uniform quality, nor is it equally interesting to different audiences. Recommender systems guide people to items they are likely to like, based on their own and other people's subjective re-into a multi-billion dollar advertising marketplace. But to the extent that people rely on recommender systems of various kinds to guide their attention, there are also natural incentives for promoters to manipulate the recommendations. An attacker may rate strategically rather than honestly and may introduce multiple entities, sometimes called sybils, to rate on behalf of the attacker.

We offer a manipulation-resistance algorithm, called the Influence Limiter, that can be overlaid on existing recommender algorithms. Consider the predictions about whether a particular target person will like various items. Each rater begins with a very low but non-zero reputation score. The current reputation limits the influence she can have on the prediction for the next item. Eventually, the target person indicates whether she likes the item and the raters who contributed to predicting whether the target would like it gain or lose reputation. The more that a rating implies a change in the prediction for the target, the greater the potential change in the rater's reputation score. A rater who simply goes along with the previous change will have no impact and thus get no change in her reputation.

The Influence Limiter has several desirable properties. First, in order to maximize the expected reputation score of a single rater endowed with some information about the target's likely response to the items, the optimal strategy is to induce predictions that accurately reveal that rater's information about the items. If the underlying recommendation algorithm is making optimal use of ratings, this implies that entering honest ratings is optimal. An important special case is that a rater who has not interacted with an item, and therefore has no information about the target's likely response to it, can only lose reputation in expectation by giving a rating.
Looking Ahead:
The Dark Side
Looking Ahead: The Dark Side

Examples:

- eBay fraud
- Fake Amazon reviews
- Gold diggers
- Griefers, trolls, sockpuppets
- Human flesh search engine
- Kasparov against the World
- Link spam
- Obama open dialog poll

- Porn monkeys
- Predicting sexual orientation from social media
- ReCAPTCHA
- Reverse engineering SSNs
- Subvert and Profit
- Trapster
- Wikipedia vandalism
The 3rd Human Computation Workshop (HCOMP 2011)

Welcome  About the Workshop  Program  Tutorial  Previous Workshops

WHERE
Co-located with AAAI 2011
San Francisco, CA

WHEN
August 8, 2011

IMPORTANT DATES
Paper Deadline: April 29, 2011
Notification: May 23, 2011
Camera Ready: May 27, 2011

SUBMISSION
Paper Format: AAAI
Submission via: CMT

WORKSHOP POSTER
to download and print!
Courtesy of Polo Chau

HCOMP 2011 has a new home!

Human Computation is the study of systems where humans perform a major part of the computation or are an integral part of the overall computational system. Over the past few years, we have observed a proliferation of related workshops, new courses, and tutorials, scattered across many conferences. In this 3rd Human Computation Workshop (HCOMP 2011), we hope to draw together participants across disciplines -- machine learning, mechanism and market design, information retrieval, decision...
Collective Intelligence and Machine Learning

Haym Hirsh
Professor, Department of Computer Science, Rutgers University
Visiting Scholar, Center for Collective Intelligence, MIT
hirsh@cs.rutgers.edu