Matt Lease

Associate Professor
School of Information
http://ir.ischool.utexas.edu
ml@utexas.edu
Talk slides: slideshare.net/mattlease

Research Areas
Information Retrieval & Search Engines
Crowdsourcing & Human Computation

- **NIST TREC 2010-2013**: Ran tracks for the US National Institute of Standards & Technology (NIST) Text REtrieval Conference (TREC)
- **Tutorials**: ACM SIGIR 2011-12 & WSDM 2011, SIAM Data Mining 2013
- **Information Retrieval & Search**
  Neural Information Retrieval: At the End of the Early Years. *IRJ* 2018
- **Human Computation & Crowdsourcing** (e.g., Human-in-the-loop)
What’s an Information School?

“The place where people & technology meet”

Wobbrock et al., 2009

“iSchools” now exist at over 100 universities around the world
Human-centered Technology Design
HCI / UX Design - undergraduate minor

HCI - Human Computer Interaction | UX – User Experience

Key Areas:

- Understanding Users
- Design & Prototyping
- Evaluation

Upper division courses

2. INF315C - Topics in HCI: User Research
3. INF350C - Advanced Topics: Designing Rich User Experiences
4. INF350C - Advanced Topics: Evaluation of Interactive Systems

5. Project / UX Capstone

Lower division courses

- INF304D - 1: Intro to Information Studies with Intro to HCI/UX
Information & Computer Science

Integrated 5-Year Degree Program

Bachelors in CS + Masters in IS
Empowering our students to be valuable contributors across the UX lifecycle

**THEORY**
- Sensation and Perception
- Decision Making
- Human Error
- Information Processing
- Mental Models
- Human Learning and Memory
- Situational Awareness

**ANALYSIS**
- Requirement Gathering
- Stakeholder Interviews
- Competitive Analysis
- Contextual Inquiry
- Survey Creation
- Interviewing
- Task Analysis

**DESIGN**
- Translating Requirements
- Journey Map
- Personas
- User Scenarios
- Information Architecture
- Prototyping
- Visual Design and Composition

**EVALUATION**
- Psychophysiological: HR, Eye Tracking, HR, GSR, EEG
- Inspection Methods: Heuristic Evaluation, Cognitive Walkthrough
- User Testing: Lab, Remote, Unmoderated, Survey and Field

**IMPLEMENTATION**
- Agile Methods
- Kanban
- Sprint
- Advocacy for UX
- Cost-Justifying / ROI
- Design Specifications
- Organizational Structures

**UX Track - Course Offerings**

<table>
<thead>
<tr>
<th>Semester 1 &amp; 2</th>
<th>Semester 3 &amp; 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human-Machine Interaction (INF 385T)</td>
<td>Design Thinking (INF 385T)</td>
</tr>
<tr>
<td>Usability (INF 385T)</td>
<td>Presenting Information (INF 385T)</td>
</tr>
<tr>
<td>Understanding Research (INF 387C)</td>
<td>Visualization (INF 385T)</td>
</tr>
<tr>
<td>Understanding and Serving Users (INF 382C)</td>
<td>Managing Information Organizations (INF 382C)</td>
</tr>
<tr>
<td>Human-Computer Interaction (INF 385C)</td>
<td>Advanced Usability (INF 385C)</td>
</tr>
<tr>
<td>Information Architecture &amp; Design (INF 385C)</td>
<td>Design Thinking II (INF 385T)</td>
</tr>
<tr>
<td>Interaction Design (INF 385T)</td>
<td>Prototyping (INF 385T)</td>
</tr>
<tr>
<td>Mobile Interaction Design (INF 385T)</td>
<td>Projects in Human-Computer Interaction (INF 385T)</td>
</tr>
</tbody>
</table>
UT Austin “Moonshot” Project

**Goal**: design a future of AI & autonomous technologies that are beneficial — not detrimental — to society.

http://goodsystems.utexas.edu
Fact Checking with Search: Misinformation & Human-AI Partnerships

Matt Lease (University of Texas at Austin)
“Truthiness” is not a new problem

“Truthiness is tearing apart our country... It used to be, everyone was entitled to their own opinion, but not their own facts. But that’s not the case anymore.”

— Stephen Colbert (Jan. 25, 2006)

“You furnish the pictures and I’ll furnish the war.”
— William Randolph Hearst (Jan. 25, 1898)
Information Literacy


“Though we may know how to find the information we need, we must also know how to evaluate it. Over the past decade, we have seen a crisis of authenticity emerge. We now live in a world where anyone can publish an opinion or perspective, true or not, and have that opinion amplified...”
Automatic Fact Checking

Exploring how artificial intelligence technologies could be leveraged to combat fake news.
Design Challenge: How to interact with ML models?

2017 ACM SIGCHI Conference on Human Factors in Computing Systems

UX Design Innovation: Challenges for Working with Machine Learning as a Design Material

Graham Dove, Kim Halskov
CAVI, Aarhus University
Aarhus, Denmark
graham.dove@cc.au.dk, halskov@cavi.au.dk

Jodi Forlizzi, John Zimmerman
HCII, Carnegie Mellon University
Pittsburgh PA, USA
forlizzi@cs.cmu.edu, johnz@cs.cmu.edu
Brief Case Study: Facebook
(simpler case: journalist fact-checking)
Tessa Lyons, a Facebook News Feed product manager: “...putting a strong image, like a red flag, next to an article may actually entrench deeply held beliefs — the opposite effect to what we intended.”
“A few classes in ‘use and users of information’ … could have helped social media platforms avoid the common pitfalls of the backfire effect in their fake news efforts and perhaps even avoided … mob rule, virality-based algorithmic prioritization in the first place.”

https://www.forbes.com/sites/kalevleetaru/
Monday, August 5, 2019
Believe it or not: Designing a Human-AI Partnership for Mixed-Initiative Fact-Checking

Joint work with
An Thanh Nguyen (UT), Byron Wallace (Northeastern), & more...

Matt Lease
School of Information
University of Texas at Austin
Want to get involved in research?

- Take an Independent Study for credit
- Find a paid Undergraduate Research Assistant job

*EUREKA*: [www.utexas.edu/research/eureka](http://www.utexas.edu/research/eureka)
Automatic Fact-Checking

Given a claim:
Facebook Shut Down an AI Experiment Because Chatbots Developed Their Own Language.

and relevant article headlines:
No, Facebook Did Not Panic and Shut Down an AI Program That Was Getting Dangerously Smart.
source: gizmodo.com

Predict headline stance: For Against Observing
Predict claim veracity: False True Unknown

Predict stance from text features (Ferreira & Vlachos 2016).
Predict veracity from stance+source features (Popat et al. 2017)
Design Challenges

• Fair, Accountable, & Transparent (AI)
  – Why trust “black box” classifier?
  – How do we reason about potential bias?
  – Do people really only want to know “fact” vs. “fake”?
  – How to integrate human knowledge/experience?
    • Joint AI + Human Reasoning, Correct Errors, Personalization

• How to design strong Human + AI Partnerships?
  – Horvitz, CHI’99: mixed-initiative design
  – Dove et al., CHI’17 “Machine Learning As a Design Material”
Nguyen et al., UIST'18

Claim Checker
Enter a Claim

Check Claim  Try A Random Claim

Example Claims:
Vice Media CEO Shane Smith paid 300,000 for a Las Vegas dinner
ISIS fighters were caught trying to enter the U.S. via the U.S.-Mexico border
A N.Y. high schooler earned 72 million in the stock market

Demo!
Web Search

Interfaces
Simple Search Interface Refinements

• For “More results” requests, stores current ranked list with the user session and displays next set in the list.
• Integrates relevance feedback interaction with “radio buttons” for “NEUTRAL,” “GOOD,” and “BAD” in HTML form.
Other Search Interface Refinements

• Highlight search terms in the displayed document.
  – Provided in cached file on Google.

• Allow for “advanced” search:
  – Phrasal search (“..”)
  – Mandatory terms (+)
  – Negated term (-)
  – Language preference
  – Reverse link
  – Date preference

• Machine translation of pages.
Web Search Example

Search suggestions
Query-biased summarization / snippet generation
Sponsored search
Search shortcuts
Vertical search (news, blog, image)
**Web Search Example**

Did you mean: **state of texas**

**TexasOnline: Official Portal of Texas**
Texas Secretary of State Esperanza Hope Andrade is reminding Texans that February 1 is the deadline to register to vote in the 2010 Primary Elections. ...

www.texasonline.com/  -  Cached  -  Similar  -  📣  📣  📣

**TExES**
Significant changes to Texas Administrative Code (TAC) §227.10 (a)(3)(C) were approved by the State Board for Educator Certification (SBEC) on October 10, ...

www.texas.ets.org/  -  Cached  -  Similar  -  📣  📣  📣

**In the state of Texas**
In the state of Texas, "highly qualified" administrators and teachers are in great demand and in short supply. According to current census information, ...

texasreview101.com/History.htm  -  Cached  -  Similar  -  📣  📣  📣

**Local business results for state of texes near Austin, TX**

- The Bob Bullock Texas State History Museum: IMAX Theatre
  www.thestoryoftexas.com - (512) 936-4639 - 50 reviews

- State of Texas: Capitol Visitors Center
  www.tspb.state.tx.us - (512) 463-0063 - 19 reviews

- State of Texas: General Information
Web Search Example

Haiti Earthquake Relief
Donate $25 to Help Children and Families Hurt by the 7.0 Earthquake.
www.worldvision.org/haiti

Haiti Earthquake Disaster
www.foodforthepoor.org

Aid Haiti Quake Victims
Habitat for Humanity is working on shelter for victims. Donate today!
www.habitat.org/Haiti-Earthquake

Haiti - Latest News
Haiti earthquake: Monday news updates
- CNN - 10 minutes ago
12:20 p.m. Monday, January 18 -- Former Senate Majority Leader Bill Frist, a medical doctor, arrived in Port-au-Prince, Haiti, on Monday to help in the relief... full story

U.S. troops boost Haiti aid security as looters swarm
Why does Haiti suffer so much? - CNN - 14 minutes ago
More US troops, UN peacekeepers expected for Haiti - AP via Yahoo! News - 15 minutes ago
more Haiti news...
Web Search Example

ALL RESULTS

Aid Haiti Quake Victims - www.Habitat.org/Haiti-Earthquake
Habitat for Humanity is working on shelter for victims. Donate today!

Haiti Earthquake Aid - www.SavetheChildren.org/donate
Millions of people affected. Donate now to help us respond.

News about Haiti

Rebuilding Haiti
Haiti’s infrastructure for things like clean water and sewage disposal was primitive before last... - World 2 hours ago
Man rescued from Haiti hotel plans to return - WTOP 44 minutes ago
French minister criticizes US aid role in Haiti - WTOP 2 hours ago

Images of Haiti

Haiti - Wikipedia, the free encyclopedia
Haiti (Haitian Creole: Ayiti), officially the Republic of Haiti (République d’Haiti; Repiblik Ayiti) is a Haitian Creole- and French-speaking Caribbean country. Along with ...
Cross-Lingual IR

• 2/3 of the Web is in English
• About 50% of Web users do not use English as their primary language
• Many (maybe most) search applications have to deal with multiple languages
  – *monolingual search*: search in one language, but with many possible languages
  – *cross-language search*: search in multiple languages at the same time
Cross-Lingual IR

Ideal

• Let user express query in native language
• Search information in multiple languages
• Translate results into user’s native language
Vertical Search

• Aka/related: federated / distributed / specialty
• Searching the “Deep” web
• One-size-fits-all vs. niche search
  – Query formulation, content, usability/presentation
Clustering Results

• Group search results into coherent “clusters”:
  – “microwave dish”
    • One group of on food recipes or cookware.
    • Another group on satellite TV reception.
  – “Austin bats”
    • One group on the local flying mammals.
    • One group on the local hockey team.

• **Northern Light** used to group results into “folders” based on a pre-established categorization of pages (like DMOZ categories).

• Alternative is to dynamically cluster search results into groups of similar documents.
Other Visual Interfaces
Speech Queries are Longer

![Graph showing the comparison between speech and text queries by word count. The graph indicates that speech queries tend to be longer than text queries, with a peak at around 4-5 words for speech compared to 2-3 words for text. The data is from Microsoft internal data April 2015.](image-url)
User Query Length

• Users tend to enter short queries.
  – Study in 1998 gave average length of 2.35 words.

• Evidence that queries are getting longer.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Jan-08</th>
<th>Dec-08</th>
<th>Jan-09</th>
<th>Year-over-year percent change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 word</td>
<td>20.96%</td>
<td>20.70%</td>
<td>20.29%</td>
<td>-3%</td>
</tr>
<tr>
<td>2 words</td>
<td>24.91%</td>
<td>24.13%</td>
<td>23.65%</td>
<td>-5%</td>
</tr>
<tr>
<td>3 words</td>
<td>22.03%</td>
<td>21.94%</td>
<td>21.92%</td>
<td>0%</td>
</tr>
<tr>
<td>4 words</td>
<td>14.54%</td>
<td>14.67%</td>
<td>14.89%</td>
<td>2%</td>
</tr>
<tr>
<td>5 words</td>
<td>8.20%</td>
<td>8.37%</td>
<td>8.68%</td>
<td>6%</td>
</tr>
<tr>
<td>6 words</td>
<td>4.32%</td>
<td>4.47%</td>
<td>4.65%</td>
<td>8%</td>
</tr>
<tr>
<td>7 words</td>
<td>2.23%</td>
<td>2.40%</td>
<td>2.49%</td>
<td>12%</td>
</tr>
<tr>
<td>8+ words</td>
<td>2.81%</td>
<td>3.31%</td>
<td>3.43%</td>
<td>22%</td>
</tr>
</tbody>
</table>

Note: Data is based on four-week rolling periods (ending Jan. 31, 2009; Dec. 27, 2008; and Jan. 26, 2008) from the Hitwise sample of 10 million U.S. Internet users.

Source: Hitwise, an Experian company
Spoken Search

Longer and more natural queries emerge given support for spoken input [Du and Crestiani’06]

See also: studies by Nick Belkin
Long / Verbose Web Queries

- User queries from
- Analysis by [Bendersky and Croft’09]

\[ MRR = \frac{1}{|Q|} \sum_{i=1}^{Q} \frac{1}{\text{rank}_i} \]
Spoken “Document” Retrieval
Query formulation reflects an ongoing dialog between users and search engines

- Users formulate queries for the search engine, based on a mental model of what it “understands”
- Search engines optimize their “understanding” for the (most frequent) submitted queries
- Individual session and long term, personal and aggregate

Result: query “language” is continually evolving
Verbosity and Complexity

• Complex information requires complex description
  – Information theory [Shannon’51]
  – Human discourse implicitly respects this [Grice’67]

• Simple searches easily expressed in keywords
  – navigation: “alaska airlines”
  – information: “american revolution”

• Verbosity naturally increases with complexity
  – More specific information needs [Phan et al.’07]
  – Iterative reformulation [Lau and Horvitz’99]
Blog Search
μ-Blog Search (e.g. Twitter)
Book Search

• Find books or more focused results
• Detect / generate / link table of contents
• Classification: detect genre (e.g. for browsing)
• Detect related books, revised editions
• Challenges
  – Variable scan quality, OCR accuracy
  – Copyright
  – Monetary model
Other IR-Related Tasks

- Automated document categorization
- Information filtering (spam filtering)
- Information routing
- Automated document clustering
- Recommending information or products
- Information extraction
- Information integration
- Question answering
## Dimensions of IR

<table>
<thead>
<tr>
<th>Content</th>
<th>Applications</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
<td>Web search</td>
<td>Ad hoc search</td>
</tr>
<tr>
<td>Images</td>
<td>Vertical search</td>
<td>Filtering</td>
</tr>
<tr>
<td>Video</td>
<td>Enterprise search</td>
<td>Classification</td>
</tr>
<tr>
<td>Scanned docs</td>
<td>Desktop search</td>
<td>Question answering</td>
</tr>
<tr>
<td>Audio</td>
<td>Forum search</td>
<td></td>
</tr>
<tr>
<td>Music</td>
<td>P2P search</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Literature search</td>
<td></td>
</tr>
</tbody>
</table>
Routing / Filtering

• Given standing query, analyze new information as it arrives
  – Input: all email, RSS feed or listserv, …
  – Typically classification rather than ranking
  – Simple example: Ham vs. spam
  – Anomaly detection
Collaborative Search
Expertise Search

There are 9 records matching your search request:

Area of Expertise = Information Retrieval

Your search took 0.250 seconds to perform.

Name: Mr John Alcock
Town/County: London
Organisation: Bristows, Solicitors
Occupation: Solicitor and European Patent/Trade Mark Attorney

Name: Mr Matthew J Altha
Town/County: Wigan, Lancs
Organisation: Independent Drug Monitoring Unit (IDMU)
Occupation: Drug Abuse Research & Information Consultant

Name: Miss Annette Clarey
Town/County: Slough, Berks
Organisation: BioMark Forensics Ltd
Occupation: Forensic Biologist

Name: Mr Andrew Fox
Town/County: Plymouth, Devon
Organisation: Audax Digital Forensic
Occupation: Computer Forensic Consultant
**Question Answering & Focused Retrieval**

**Question:**

what is information retrieval

**Possible answers:**

1. There is a common confusion, however, between data retrieval, document retrieval, information retrieval, and text retrieval, and each of these is its own bodies of literature, theory, praxis and technologies.
2. Information Retrieval (IR) is a discipline of studying theories, models, and techniques that deal with the representation, storage, organization, an retrieval of information items so that they can be useful to humans.
Community QA

Yahoo! Answers

News & Events

Did bill clinton really engineer this economic crisis?
1 In Current Events - Asked by eja123 - 7 answers - 45 minutes ago

Why did Bill o'reilly quit his daily radio show?
1 In Media & Journalism - Asked by davidagoldsmith - 1 answer - 5 minutes ago

Why do people think Obama won cuz he's black?
1 In Current Events - Asked by jade3712 - 7 answers - 3 minutes ago

How many people did Saddam Hussein kill?
1 In Current Events - Asked by eja123 - 2 answers - 10 minutes ago

Is obama going to improve foreign relations?
1 In Current Events - Asked by eja123 - 4 answers - 10 minutes ago

Do You Think Chris McCandless "Deserved" to Die?
1 In Media & Journalism - Asked by veggieessi - 1 answer - 7 minutes ago

Obama haters, were you satisfied with the job the Obama administration did the past 8 years?
1 In Current Events - Asked by Mallory - 5 answers - 6 minutes ago

What qualities should a Journalist have?
1 In Media & Journalism - Asked by suzanne - 3 answers - 9 minutes ago

WikiAnswers.com

Politics and Society

Why is Latin America called 'Latin' America? [Edit categories]
Popularity: 335

What is the difference between Saudi Arabia and Arabia?
[Edit categories]
Popularity: 75

What were the Maya achievements?
[Edit categories]
Popularity: 48

Wondir

Science

11 Mar '09, 00:56 (SCI) What is the scientific name for dragonfly ants? (1 response)
10 Mar '09, 14:40 (SCI) what are three uses of boron chemicals (1 response)
10 Mar '09, 09:12 (SCI) Does nucleus only contain protons and neutrons? (2 responses)
10 Mar '09, 02:16 (SCI) what colour is violet???????? (3 responses)
10 Mar '09, 02:00 (SCI) how has globalisation had an impact on the role of the state and the concept of sovereignty
10 Mar '09, 01:48 (SCI) is s. equi sub specie equi an aerobic bacteria?
10 Mar '09, 00:01 (SCI) when did the sugar act occur? (1 response)
09 Mar '09, 16:16 (SCI) Does it rain sperm at a baby shower? (3 responses)
09 Mar '09, 09:30 (SCI) How much stronger is the new vitreous carbon material invented by the Tokyo Institute of Technology compared with the material made from cellulose?
08 Mar '09, 22:36 (SCI) where can i go to find answers about alcohol related questions? (2 responses)
07 Mar '09, 22:43 (SCI) if the earth suddenly stopped spinning ,would we go flying through the air? (4 responses)
e-Discovery

Electronic Discovery Reference Model
www.edrm.net

Electronic Discovery Lessons from Dedicated Review Teams

Podcast: Play in new window | Download

Listen to Matt Clarke, a shareholder and member of the Riley Carlock & Applewhite’s Document Control Group, and Karl Schleneman, Director of Legal Analytics & Review at Juttinek as moderator, as they discuss working with dedicated review spaces for staffing electronic discovery projects. Matt is the
Systematic Review is e-Discovery in Doctor’s Clothing

Joint work with

<table>
<thead>
<tr>
<th>Gordon V. Cormack (U. Waterloo)</th>
<th>An Thanh Nguyen (U. Texas)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thomas A. Trikalinos (Brown U.)</td>
<td>Byron C. Wallace (U. Texas)</td>
</tr>
</tbody>
</table>

SIGIR 2016 Workshop on Medical IR (MedIR)
Hybrid Man-Machine Relevance Judging

• Systematic review (medicine) and e-Discovery (law / civil procedure) have traditionally relied on trusted doctors/lawyers for judging

• Automatic relevance classification is more efficient but less accurate

• Recent active learning work has investigated hybrid man-machine judging combinations
  – e.g., TAR & TREC Legal Track, recent CLEF track
What is Active Learning?

(Wallace et al., 2016)
Information Retrieval and Web Search

Introduction
Relevance

• Relevance is a subjective judgment and may include:
  – Being on the proper subject.
  – Being timely (recent information).
  – Being authoritative (from a trusted source).
  – Satisfying the goals of the user and his/her intended use of the information (*information need*).
Relevance

• What is it?
  – **Simplistic definition**: A relevant document contains the information that a person was looking for when they submitted a query to the search engine
  – Many factors influence a person’s decision about what is relevant: e.g., task, context, novelty, style
  – *Topical relevance* vs. *user relevance*
Who and Where?
Modeling Relevance

- **Ranking algorithms** used in search engines
- Ranking is typically statistical and based on its *observable* properties rather than *underlying* linguistic properties
  - i.e. counting simple text features such as words instead of inferring underlying linguistic syntax
  - However, both kinds of *features / evidence* can be incorporated into a statistical model
Keyword Search

• Simplest notion of relevance is that the query string appears verbatim in the document.

• Slightly less strict notion is that the words in the query appear frequently in the document, in any order (\textit{bag of words}).
Problems with Keywords

• May not retrieve relevant documents that include synonymous terms.
  – “restaurant” vs. “café”
  – “PRC” vs. “China”

• May retrieve irrelevant documents that include ambiguous terms.
  – “bat” (baseball vs. mammal)
  – “Apple” (company vs. fruit)
  – “bit” (unit of data vs. act of eating)
Users and Information Needs

• Search evaluation is user-centered

• Keyword queries are often poor descriptions of actual information needs

• Interaction and context are important for inferring user intent

• Query refinement techniques such as query expansion, query suggestion, relevance feedback improve ranking
Query Disambiguation

• Given (typically terse like “apple”) query, infer possible underlying intents / needs / tasks

• With longer queries, detect key concepts and/or segment (e.g. “new york times square”)
More Applications…
Location-based Search
See something you want to remember
When you notice an item to remember, tap “Remembers” in the Amazon App.

Snap photo & send
Your iPhone camera will open. Take a photo of the item and it will be sent to Amazon.

See reminders
Your photos & any similar products that Amazon finds are stored in the app and on Amazon.com.
Content-based music search
Retrieving Information, not Documents

**Events related to "haiti"**

**Timeline**

<table>
<thead>
<tr>
<th>Year</th>
<th>Event Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1804</td>
<td>Haiti attempted to establish closer ties with the United States</td>
</tr>
<tr>
<td>1804</td>
<td>January 1, 1804: It is also known as Haiti's independence day</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>Event Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1793</td>
<td></td>
</tr>
<tr>
<td>1794</td>
<td></td>
</tr>
<tr>
<td>1795</td>
<td></td>
</tr>
<tr>
<td>1796</td>
<td></td>
</tr>
<tr>
<td>1797</td>
<td></td>
</tr>
<tr>
<td>1798</td>
<td></td>
</tr>
<tr>
<td>1799</td>
<td></td>
</tr>
<tr>
<td>1800</td>
<td></td>
</tr>
<tr>
<td>1801</td>
<td></td>
</tr>
<tr>
<td>1802</td>
<td></td>
</tr>
<tr>
<td>1803</td>
<td></td>
</tr>
<tr>
<td>1804</td>
<td></td>
</tr>
<tr>
<td>1805</td>
<td></td>
</tr>
<tr>
<td>1806</td>
<td></td>
</tr>
<tr>
<td>1807</td>
<td></td>
</tr>
<tr>
<td>1808</td>
<td></td>
</tr>
<tr>
<td>1809</td>
<td></td>
</tr>
<tr>
<td>1810</td>
<td></td>
</tr>
<tr>
<td>1811</td>
<td></td>
</tr>
<tr>
<td>1812</td>
<td></td>
</tr>
<tr>
<td>1813</td>
<td></td>
</tr>
<tr>
<td>1814</td>
<td></td>
</tr>
</tbody>
</table>

**Events in the timeline**

**January 1, 1804**

(From W.Gonaives) "It is also known as Haiti's independence day because it was there that Gen. Jean-Jacques Dessalines declared Haiti's independence on January 1, 1804."

(From W.Gonaives) "Gonaives is also known as Haiti's city of Independence because it was there that Jean-Jacques Dessalines declared Haiti, the former Saint-Domingue, independent from France on January 1, 1804 by reading the Act of Independence, drafted by Boisrond Tonnerre, on the Place d'Armes of the town."

(From W.List of French Governors of Saint-Domingue) "January 1, 1804, Independence of Haiti, Jean-Jacques Dessalines is Provisional Chief of the Haitian Government to September 22, 1804 and then Emperor of Haiti until October 17, 1806."
News Tracking (Living Stories)

Washington Tackles Health Care Reform

Overview: The House and Senate have approved sweeping legislation that would provide health care insurance for most Americans, at huge cost to the government. The House plan, approved Nov. 7 in an almost strictly party-line vote, would spend $1.06 trillion to extend coverage to about 36 million Americans. The Senate bill, passed on Dec. 24, went through several iterations before attracting the support of a filibuster-proof coalition of 60 votes. It would cost $871 billion and give coverage to 31 million people who lack it now.

Read more...

Timeline of important events

- White House, unions reach deal on 'Cadillac' tax
  Jan 14, 2010

- Senate approves landmark health-care bill
  Dec 24, 2009

- Health care bill clears last Senate hurdle before passage
  Dec 23, 2009

- Senate health-care bill clears crucial hurdle in 60-40 vote
  Dec 21, 2009

- Senate reaches deal on health-care reform
  Dec 20, 2009

- White House, unions reach deal on 'Cadillac' tax
  Jan 14, 2010

- Senate approves landmark health-care bill
  Dec 24, 2009

- Health care bill clears last Senate hurdle before passage
  Dec 23, 2009

- Senate health-care bill clears crucial hurdle in 60-40 vote
  Dec 21, 2009

- Senate reaches deal on health-care reform
  Dec 20, 2009

Related

- Obama pleads for pragmatism on health-care overhaul - Feature

Read more...

- White House, unions reach deal on 'Cadillac' tax
  Jan 14, 2010

A bargain on taxing high cost health insurance policies puts negotiators close to an overall deal on a health-care reform plan. Under the agreement, family plans that cost more than $24,000 and individual policies that cost more than $9,900 would be subject to a 40 percent surtax. The tax would be imposed on the insurance company, but economists believe it would be passed on to workers.

Last year, the average family policy in America cost $13,375, according to a survey by the Kaiser Family Foundation.

The changes would cut revenue for the health reform package by $60 million over 10 years, a sum likely to be made up by...
Memetracker
“Hyper-local” Search

EveryBlock
A news feed for your block

Track and discuss what’s new in your neighborhood.

Choose a city...

- Atlanta
- Boston
- Charlotte
- Chicago
- Dallas
- Detroit
- Houston
- Los Angeles
- Miami
- New York
- Philadelphia
- San Francisco
- San Jose
- Seattle
- Washington, DC

Don’t see your city?

Find news near you
Enter an address and see nearby crime, news coverage, neighbor announcements and more.

Browse news by topic
Get a citywide overview of news by category, then click on topics to find what you’re looking for.

Explore your city
Choose a neighborhood or ZIP code and see what’s happening nearby or track trends over time.

Check out the free EveryBlock iPhone app.
Recent IR History

• **2010’s**
  - Intelligent Personal Assistants
    - Siri
    - Cortana
    - Google Now
    - Alexa
  - Complex Question Answering
    - IBM Watson
  - Distributional Semantics
  - Deep Learning
Deep (a.k.a. Neural) IR
Growing Interest in “Deep” IR

• **Success of Deep Learning (DL) in other fields**
  – Speech recognition, computer vision, & NLP

• **Growing presence of DL in IR research**
  – e.g., SIGIR 2016 Keynote, Tutorial, & Workshop

• **Adoption by industry**
  – WIRED: [AI is Transforming Google Search](https://www.wired.com/2016/02/algorithms-affect-google-search/). The Rest of the Web is next. February, 2016.

[https://en.wikipedia.org/wiki/RankBrain](https://en.wikipedia.org/wiki/RankBrain)
But Does IR Need Deep Learning?

• Chris Manning (Stanford)’s SIGIR Keynote: “I’m certain that deep learning will come to dominate SIGIR over the next couple of years... just like speech, vision, and NLP before it.”

• Despite great successes on short texts, longer texts typical of ad-hoc search remain more problematic, with only recent success (e.g., Guo et al., 2016)

• As Hang Li eloquently put it, “Does IR (Really) Need Deep Learning?” (SIGIR 2016 Neu-IR workshop)
Neural Information Retrieval: A Literature Review

Figure 4: Two basic neural architectures for scoring the relevance of queries to documents.

Ye Zhang et al.  
https://arxiv.org/abs/1611.06792  
*Posted 18 November, 2016*
Word Embeddings
Traditional “one-hot” word encoding

Leads to famous *term mismatch* problem in IR

The standard word representation

The vast majority of rule-based and statistical NLP work regards words as atomic symbols: hotel, conference, walk

In vector space terms, this is a vector with one 1 and a lot of zeroes

\[
[0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0]
\]

Dimensionality: 20K (speech) – 50K (PTB) – 500K (big vocab) – 13M (Google 1T)

We call this a “one-hot” representation. Its problem:

\[
\text{motel} \ [0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0] \ \text{AND} \ \\
\text{hotel} \ [0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0] = 0
\]
Distributional Representations

Define words by their co-occurrence signatures

You can get a lot of value by representing a word by means of its neighbors

“You shall know a word by the company it keeps”
(J. R. Firth 1957: 11)

One of the most successful ideas of modern statistical NLP

government debt problems turning into banking crises as has happened in
saying that Europe needs unified banking regulation to replace the hodgepodge

These words will represent \textit{banking}

You can vary whether you use local or large context
to get a more syntactic or semantic clustering
“Early” Neural Word Embeddings

- word2vec (Mikolov et al., 2013) – sliding window
  - CBOW: predict center word given window context
  - Skip-gram: predict context given center word

- See also: GloVe (Pennington et al., 2014)
Extending IR Models with Word Embeddings
## Recent IR Work with Word Embeddings

<table>
<thead>
<tr>
<th>Task</th>
<th>Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bug Localization</td>
<td>Ye et al. (2016)</td>
</tr>
<tr>
<td>Contextual Suggestion</td>
<td>Manotumruksa et al. (2016)</td>
</tr>
<tr>
<td>Cross-lingual IR</td>
<td>BWESG (Vulic and Moens (2015))</td>
</tr>
<tr>
<td>Detecting Text Reuse</td>
<td>Zhang et al. (2014)</td>
</tr>
<tr>
<td>Domain-specific</td>
<td>De Vine et al. (2014)</td>
</tr>
<tr>
<td>Semantic Similarity</td>
<td></td>
</tr>
<tr>
<td>Community Question</td>
<td>Zhou et al. (2015)</td>
</tr>
<tr>
<td>Answering</td>
<td></td>
</tr>
<tr>
<td>Short Text Similarity</td>
<td>Kenter and de Rijke (2015)</td>
</tr>
<tr>
<td>Outlier Detection</td>
<td>ParagraphVector (Le and Mikolov (2014))</td>
</tr>
<tr>
<td>Sponsored Search</td>
<td>Grbovic et al. (2015b), Grbovic et al., (2015a)</td>
</tr>
</tbody>
</table>
Ponte & Croft (2001): LM for IR

\[
P(D|Q) = \left[ P(Q|D) \ P(D) \right] / P(Q)
\]

\(\propto P(Q|D) \ P(D)\) for fixed query

\(\propto P(Q|D)\) assume uniform \(P(D)\)

\[
P(Q|D) = \prod_q \alpha * P(q|D) + (1 - \alpha)P(q|C)
\]
Berger & Lafferty (1999)

• IR as Statistical Translation
  – Document $d$ contains word $w$
  – $w$ is translated to observed query word $q$

$$p_\alpha(q \mid d) = \alpha p(q \mid D) + (1 - \alpha) p(q \mid d)$$

$$= \alpha p(q \mid D) + (1 - \alpha) \sum_{w \in d} l(w \mid d) t(q \mid w)$$
GLM: Ganguly et al., SIGIR 2015
GLM: Ganguly et al., SIGIR 2015

\[ P(t|t', d) = \frac{\text{sim}(t, t')}{\sum_{t'' \in d} \text{sim}(t, t'')} = \frac{\text{sim}(t, t')}{\Sigma(d)} \]

\( \text{sim}(t, t') \) is the cosine similarity between the vector representations of \( t \) and \( t' \) and \( \Sigma(d) \) is the sum of the similarity values between all term pairs.
and Lafferty have proposed an alternative estimation of $p_s(w|d)$ inspired by models in statistical machine translation [5].

$$p_t(w|d) = \sum_{u \in d} p_t(w|u)p(u|d)$$

**Estimating Translation Probabilities with Neural Language Models**

The use of neural language models based on continuous bag-of-words or skipgram gives rise to two different word embeddings. Word embeddings can be used to estimate translation probabilities in translation language models; specifically, cosine similarity can be used as a proxy for $p(u|w)$:

$$p_{cos}(u|w) = \frac{\cos(u, w)}{\sum_{w' \in V} \cos(u', w)}$$
**DeepTR**: Zheng & Callan, SIGIR 2015

- Supervised learning of effective term weights
  - Like *RegressionRank* (Lease et al., ECIR 2009), (Lease, SIGIR 2009) but without feature engineering
- Represent each query term in context by avg. query embedding - term embedding

**DeepTR-BOW:**

\[
\#weight( \hat{P}(\text{apple}|R) \text{ apple} \\
\hat{P}(\text{pie}|R) \text{ pie} \\
\hat{P}(\text{recipe}|R) \text{ recipe} )
\]

**DeepTR-SD:**

\[
\#weight( \\
0.8 \#weight( \hat{P}(\text{apple}|R) \text{ apple} \\
\hat{P}(\text{pie}|R) \text{ pie} \\
\hat{P}(\text{recipe}|R) \text{ recipe} ) \\
0.1 \#combine(\#1(\text{apple pie}) \\
\#1(\text{pie recipe}) ) \\
0.1 \#combine( \\
\#uw8(\text{apple pie}) \\
\#uw8(\text{pie recipe}) )
\]
• Learn topical word embeddings *at query-time*
  – New flavor of classic IR *global vs. local* tradeoff
  – Compare use of collection vs. external corpora
• No comparison to pseudo-relevance feedback

Figure 5: Global versus local embedding of highly relevant terms. Each point represents a candidate expansion term.
Cross-Lingual IR with Bilingual Word Embeddings
Bilingual Word Embeddings for Phrase-Based Machine Translation

Will Y. Zou†, Richard Socher, Daniel Cer, Christopher D. Manning
Department of Electrical Engineering† and Computer Science Department
Stanford University, Stanford, CA 94305, USA

the cat sat on the mat

on the mat sat the cat

course the house is small

natürliche ist das haus klein

Australia is one of a few countries that has relationship with North Korea
BilBOWA: Fast Bilingual Distributed Representations without Word Alignments

Stephan Gouws, Yoshua Bengio, Gregory S. Corrado - ICML - 2015
Merge & Shuffle: Training a SGNS (or any other monolingual model!) on shuffled “pseudo-bilingual” documents
Ye et al., ICSE 2016: Finding Bugs

• Given textual bug report (query), find software files needing to be fixed (documents)
  – Saha, Lease, Khursid, Perry (ASE, 2013)

• Augment the Skip-gram model to predict all code tokens from each text word, and all text words from each code

Figure 9: Example of semantically related text and code, from API documents.

Figure 10: Positive pairs generated from semantically related text and code.
Going Deeper with Characters

“...The dominant approach for many NLP tasks are recurrent neural networks, in particular LSTMs, and convolutional neural networks. However, these architectures are rather shallow in comparison to the deep convolutional networks which are very successful in computer vision.

We present a new architecture for text processing which operates directly on the character level and uses only small convolutions and pooling operations. We are able to show that the performance of this model increases with the depth: using up to 29 convolutional layers, we report significant improvements over the state-of-the-art on several public text classification tasks. To the best of our knowledge, this is the first time that very deep convolutional nets have been applied to NLP.”
Resources

http://deeplearning.net
# Neural IR Source Code Released

<table>
<thead>
<tr>
<th>System</th>
<th>Citation</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>word2vec</td>
<td>Mikolov and Dean (2013)</td>
<td><a href="https://code.google.com/archive/p/word2vec/">https://code.google.com/archive/p/word2vec/</a></td>
</tr>
<tr>
<td>GloVe</td>
<td>Pennington et al. (2014)</td>
<td><a href="http://nlp.stanford.edu/projects/glove/">http://nlp.stanford.edu/projects/glove/</a></td>
</tr>
</tbody>
</table>
Thank You!

Slides: slideshare.net/mattlease
Lab: ir.ischool.utexas.edu