E6895 Advanced Big Data Analytics Lecture 10:

Social and Cognitive Analytics

Ching-Yung Lin, Ph.D.

Adjunct Professor, Dept. of Electrical Engineering and Computer Science
Combining Signal Processing, Data & Graph Mining, and Machine Learning by multi-modality signal sensing & understanding.
People Analytics ➔ All Aspects of a Person

Valuable Source Code

Access Logs
- Who (accounts)
- When
- From where ...

E-Mail, IM, Web,
- A critical foundation to:
  - Anomaly Detection at Multiple Scales
  - Information Leakage Detection
  - Real-Time Intelligence Filtering

Device Traffic

Internal Employee Data

Laptop

Mobile

Professional / Social Networking

Friends / Network

Complete Data of Logs, Documents, Slides, Emails, IM, Clicks, etc.

Professional Network, Twitter Network regarding to work
Example — our enterprise social analytics system

- 15,000 volunteers; 76 countries; 119,000 users
- 25,000,000 emails & instant messages (incl. content)
- 1,500,000 Learning click data; 44,000 entities
- 6,681,000 Knowledge & Sales access data; 240,000 entities
- 1,687,000 Media Library access data; 105,000 entities
- 700,000 posts (blogs 3,000, file sharing 210,000, bookmark 450,000, Wiki 11,000) data
- 200,000 people’s consulting financial databases
- 400,000 organization/demographic data
- 100,000 intranet searches per day
Privacy – adaptive features for global privacy laws

- Privacy features and worked with GBS to go through 2-year global privacy review with privacy officers and labor union approval to make SmallBlue a deployable production system
- A unique large-scale social network capturing and process system that is lawful & user-aware system about people worldwide
Social Mining Enterprise Architecture

- novel distributed data collection,
- novel expertise inference algorithm,
- novel social network inference,
- novel visualization and analysis
SmallBlue Enterprise 1.1 (outdated drawing; current version 2.4)

System Overview

IBM Internal Public Data
- BluePages Server
- User Register
- Non-Searchable Terms
- Knowledge Keyword Extractor
- BPS Info Indexer
- Find Search Engine
- SmallBlue Find Components
- SmallBlue Net Components

Contributors
- Weekly Index creation / update
- Keyphrase, graph link info

Knowledge Interface Engine
- Search Engine
- Knowledge Keyword Indexer
- Social Path Indexer

SmallBlue Components
- Reach Server
- DogeMap
- Blogs, Forums, Community Map

Data Store
- Communication Indexes

Ego Server
- Ego-Net Inference Engine
- Ego-Net Hygist
- SmallBlue Ego Components

Contributors, Ego-Net Hygist
- Weekly Index creation / update
- Modelled topics or all Contributors

Data Flow
- Data sent once a week / month based on Contributors SBC settings

User Application
- Data not accessible by any user or interface

SmallBlue Ego
- Knowledge Interface Engine

SmallBlue Component
- Data not accessible by any user or interface
Advanced Social Mining Enterprise 1.1 architecture

Note that this diagram indicates the exact number of machines considered necessary in each location type. The community map service will be added in future.
My personal social network automatically found by SmallBlue with social distance

**Personal Social Network Capital**

- The number of unique people that are being contributed to your extended social network by this person is 1175.
- Adding a person in personal network (i.e., frequent communications), increases $948 yearly revenue for IBM. *(selected by BusinessWeek Magazine as the Top Story of the Week, April 10, 2009)*
- 1% increase in social network diversity is associated with $239.5 in monthly revenue
- 1% increase in social network diversity is associated with an increase of 11.8% in job retention.

It can also show the evolution of my social network.

How many people in my personal networks?

What types of unique colleagues my friend Chris can help me connect to?
How are company’s employees communicating ‘healthcare’ linking with each other? Who are the key bridges? Who have the most connections? How do these people cluster? It can be extended to analyze relationship of customers.
Social Network Analysis (cont’d)

IBM Healthcare-related employees in the world

Connections between different divisions

IBM Healthcare-related employees in the U.S.

Key social bridges
Finding expertise

E.g.: Search for the most knowledgeable colleagues within my 3-degree network for who knows ‘healthcare’. (or within a country, a division, a job role, or any group/community)

As a user, you can only see their public information. Private info is used internally to rank expertise but private data can never be exposed.

Click a name to see their profile (SmallBlue Reach)

My shortest path to Susan

As a user, you can only see their public information. Private info is used internally to rank expertise but private data can never be exposed.

Click a name to see their profile (SmallBlue Reach)
Social Paths

- Is Tom a right person to me? Shortest Social Paths to any person within 6-degrees.

My various paths to Tom. SmallBlue can show the paths to any colleagues up to 6-degree away.
Topological point of views
- What type of network structure is beneficial?

**Cohesive Network**
- Trust
- Absorptive capacity
- Precision, Reliability

**Structurally Diverse Network**
- Brokering position
- Access to many pools of diverse, novel information

What type of network structure is most beneficial in a electronic network for consultants?
- Importance of Direct Contacts?
- Importance of Indirect Contacts?
- Constrained vs. unconstrained?
## Network Topology Measures

<table>
<thead>
<tr>
<th>Direct Contacts</th>
<th>Indirect Contacts</th>
<th>Structural Diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Size(7) = 4</strong></td>
<td><strong>Btw(7) = 33</strong></td>
<td><strong>Div(7) = .53</strong></td>
</tr>
<tr>
<td><strong>Size(12) = 3</strong></td>
<td><strong>Btw(12) = 6</strong></td>
<td><strong>Div(12) = 0.16</strong></td>
</tr>
<tr>
<td><strong>+ No information distortion</strong></td>
<td><strong>+ Access diverse information</strong></td>
<td><strong>+Transfer complex knowledge</strong></td>
</tr>
<tr>
<td><strong>- High maintenance cost</strong></td>
<td><strong>- Information distortion</strong></td>
<td><strong>- Access diverse knowledge</strong></td>
</tr>
<tr>
<td><strong>Network size → strong work performance (?)</strong></td>
<td><strong>Btw-centrality → Strong work performance (?)</strong></td>
<td><strong>Diversity → Strong work performance (?)</strong></td>
</tr>
</tbody>
</table>
Enterprise becomes more successful utilizing Social Network Analysis

MIT studied 2,038 IBM Global Business Consultants for 2 years, it was found that:

- After a consultant started using SmallBlue, his social network/capital obviously grew and his monthly billable revenue for IBM increased by $584.15 (i.e., $7,010 per year)

Joint analysis of social capital and economic capital:

- Adding a person in personal network (i.e., someone with frequent communications), increases $948 yearly revenue for IBM. (selected by BusinessWeek Magazine as the Top Story of the Week, April 8, 2009)
- 1% increase in social network diversity is associated with $239.5 in monthly revenue (i.e., $2,874 revenue increase per year).
- 1% increase in social network diversity is associated with an increase of 11.8% in job retention (i.e., surviving layoff).

IBM Research Achievement – “SmallBlue made Millions of Contribution to GBS in 2009”

SmallBlue / Atlas was featured in 120+ news articles, including 4 times by BusinessWeek (Jan and May 2008, April and June 2009)
Observations from Personal Social Networks vs. Revenue

- Structural Diverse networks with abundance of structural holes are associated with higher performance.
  - *Having diverse friends helps.*

- Betweenness is negatively correlated.
  - *Being a bridge between a lot of people is not helpful.*

- Network reach are highly corrected.
  - *The number of people reachable in 3 steps is positively correlated with higher performance.*

- Having too many strong links — the same set of people one communicates frequently is negatively correlated with performance.
  - *Perhaps frequent communication to the same person may imply redundant information exchange.*

  - Future textual analysis can be done to confirm this.
Project Team Composition—Managers

The number of managers in a project exhibit an inverted-U shaped curve.

1. Having managers in a project is correlated with team performance initially.
2. Too many managers in a project is negatively associated with team performance.

\[
\text{revenue} = \alpha + \beta_1 \cdot \text{mgr} + \beta_2 \cdot \text{mgr}^2 + \gamma_1 \cdot \text{otherfactor}_1 + \ldots + \gamma_k \cdot \text{otherfactor}_k + \varepsilon
\]

\[\begin{align*}
\beta_1 & = 2733.9^{***} \\
& \quad \text{(537.5)} \\
\beta_2 & = -682.02^{***} \\
& \quad \text{(215.3)}
\end{align*}\]

\[S = .027 \quad S = -.056\]
Culture Factor in CMC-based Communications

Collaborating Globally:

preferences of CMC tools

patterns of growing social network

sentiments in conversations
Preferences of CMC Tools

IM vs. Email

Calendar Meet vs. IM

JP  UK  US  CA  DE  BR  IN  CN

only IM  both IM&EM  only EM

only CA  both CA&IM  only IM
Growing one’s Social Networks

![Graph showing the growth of social networks over time. The graph plots the number of gained friends against time (by month). Different lines represent countries: US (blue), DE (dashed black), CN (red), IN (green), and JP (gray). The graph shows that the US has the highest number of gained friends, followed by DE, CN, IN, and JP.]
Sentiments in Conversation

The diagram illustrates the relationship between the degree of positive sentiments and the degree of negative sentiments for different countries. The countries are represented as points on the graph:

- **United States (US)**
- **Canada (CA)**
- **United Kingdom (UK)**
- **India (IN)**
- **China (CN)**
- **Japan (JP)**
- **Brazil (BR)**
- **Germany (DE)**

The graph shows a gradient that suggests a correlation between the degree of positive and negative sentiments, with countries like the United States and United Kingdom having a higher degree of positive sentiments compared to others.
Role Analysis

Role difference of normal behavior
Information Reuse Behavior (CHI ’11)

Percentage of slides with reused content

Number of slide pairs with exact vs. partial text reuse

Percentage of reused slides that were reused by the same author vs. by a different author

Percentage of downloaded material being reused
Behavior Detection

- Overall Flowchart: Network Science + Machine Learning + Role Mining + Visualization
Anomaly Detection – algorithms and infrastructure

▪ **Thrust 1**: Anomaly Detection Algorithms
  -- New algorithms to detect abnormal humans (nodes) as well as abnormal contacts (edges) from social networks.
  -- Explore the structure feature and incorporate content (semantic) features.

▪ **Thrust 2**: Anomaly Usability
  -- Address the ‘lack-of-the ground-truth’ issue by
    (1) Interpretation friendly properties (e.g., non-negativity, sparseness, etc) into the current anomaly detection matrix factorization; and
    (2) providing some concise summarization to perform anomaly attribution.

▪ **Thrust 3**: Infrastructure Support
  -- General and scalable graph/network management system to process large network data, especially social and behavior of people / unique IPs.
A pilot project was done by CRL in a telecomm area of 6 million users in 2009.

In experiment

- Social Network Analysis is with recall of 89.97% and precision of 88.17% while comparison system is with 66.77% recall and 14.85% precision.
- SNA’s precision/recall area is 8 times larger
Anomaly Detection – information flow-based approach

An illustrative example of an information spreading tree. This tree is of size 8, width 4, depth 3.

Probability ratio of email forwarding as a function of (a) hierarchical level difference and (b) organizational distance between initiators and spreaders. The information spreading exhibits some homophily effect.

Video demo: http://smallblue.research.ibm.com/demos/
Large-Scale Graph Analysis

- T 1.1: Structure Feature Extraction
- T 1.2: Initial Filtering
- T 1.3: Scalability
Time-to-Time (T3) Proximity Matrix Analysis

<table>
<thead>
<tr>
<th>Time Steps</th>
<th>Events</th>
<th>Entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>t₁</td>
<td>e₁</td>
<td>b₁, b₂</td>
</tr>
<tr>
<td>t₁</td>
<td>e₂</td>
<td>b₂, b₃</td>
</tr>
<tr>
<td>t₂</td>
<td>e₃</td>
<td>b₂, b₃</td>
</tr>
<tr>
<td>t₂</td>
<td>e₄</td>
<td>b₃, b₄</td>
</tr>
<tr>
<td>t₃</td>
<td>e₅</td>
<td>b₄, b₅</td>
</tr>
<tr>
<td>t₄</td>
<td>e₆</td>
<td>b₅, b₆</td>
</tr>
<tr>
<td>t₄</td>
<td>e₇</td>
<td>b₆, b₇</td>
</tr>
<tr>
<td>t₅</td>
<td>e₈</td>
<td>b₆, b₇</td>
</tr>
<tr>
<td>t₆</td>
<td>e₉</td>
<td>b₇, b₈</td>
</tr>
</tbody>
</table>

(a) Input of T3  
(b) Graph Representation  
(c) Output of T3 (at finest level)
Thrust 2: Semantic Level Analysis

- T 2.1: Semantic Feature Extraction
- T 2.2: Feature Augmentation and Selection
- T 2.3: Abnormal Event Extraction
- T 2.4: Role Mining and Analysis

Input: free text

I’m going to **vacation** in **Iran** with my sister in **January**

Attribute Name: vacation, Value: Iran, January

Attribute Name: vacation, Value: (Iran, Jan)

Attribute Name: vacation, Value: (Iran, Jun)

....

Attribute Name: vacation, Value: (Iran, XX)

Confirmed events

Event Knowledge Base

1. Event, Time & Location Identification
   (Semi-supervised learning)

2. Event - (Time, Location) Pair Assignment
   (Statistical and grammatical association)

3. Event Clustering
   (Hierarchical clustering)

4. Active learning feedback loop
   (Active Learning)

Output: Extracted Event-Attribute Pairs
Thrust 3: Ranking and Aggregation

- T 3.3: Temporal Dependency Analysis
- T 3.4: Anomaly Aggregation

T 3.1: User Behavior Modeling
T 3.2: Hierarchical Event Analysis

(a) Combined Composite Activity Model
(b) Unrolled Composite Activity Model
Example 1: Internet Map
Nodes: ISPs; Edges: Connection
(33K Nodes, 290K edges)

Example 2: Social Network
Nodes: People; Edges: Friendship
(FaceBook has 500M+ Users)

Example 3: Web Graph
Nodes: Web Pages; Edges: Hyperlinks
(Yahoo Web: 1.4B nodes, 6.6B edges)

Multiple Scales, Multiple Disciplines
Network Analysis Example: Centrality Ranking in Large Networks

"Who are the most important actors?"

**Three centralities**
- **Degree**: # of neighbors
- **Closeness**: avg. shortest path length
- **Betweenness**: # of times a node sits between shortest path

**Application**
- Measuring the financial company value
- Network attack monitoring

[15th Century Florentine Family]

- $|V| = 15$
- $|E| = 19$
- **Degree**: Easy
- **Closeness**: Easy
- **Betweenness**: Easy

$|V| = \text{Billions}$
$|E| = \text{Billions}$

- **Degree**: Easy
- **Closeness**: Hard
- **Betweenness**: Hard

$O(|E|)$
$O(|V|^3)$
$O(|V|^{2\log|V|})$

**For 2 Billion Edges,**
- standard closeness: 30,000 years
Example -- we proposed two new centralities (‘effective closeness’ and ‘LineRank’), and efficient large scale algorithms for billion-scale graphs.

Scalability Results
(Near-linear scalability)

For 2 Billion Edges,
- standard closeness: 30,000 years
- effective closeness: ~ 1 day!
1,000,000 times faster!

Analysis of Real-World Graph
Privacy – cryptography and key management approach (CIKM ’11)

- A novel methodology & system for data mining and content/people recommendations
- New cryptographic method:
  - Polynomial Ring Homomorphism, derived from Lattice-Based Cryptography
- Encrypted domain:
  - Addition
  - Multiplication
  - Division
- Key management protocol for:
  - Encrypted Multi-Layer Ranking

Diagram with nodes and arrows representing users and items, and equations for soft clustering for recommendations.
Social Network Analysis for Marketing and Sales

▪ What is the problem?
  – Past studies showed significant success on utilizing social relationships for sales & marketing.
    • McKinsey (2009) surveyed 190+ firms in all industry sectors utilizing ‘social selling’:
      -- the transaction increase consideration by avg. 19%;
      -- the average yield increase conversion by avg. 17%.
    • Krackhardt (Carnegie Mellon U, 2005) showed that companies with strong informal networks perform
      5 or 6 times better than those with weak networks.
    • Brydon (VisiblePath, 2006) showed that the performance gains of companies utilizing relationships
      are 16x in sales; 4x in marketing; and 10x in hiring
  – How to utilize Social Network Analysis for Marketing and Sales?

▪ What is the solution?
  – Conduct social graph analysis, human capital analysis, and economic analysis to quantify micro- and
    macro- social capital of each company (B2B) or each individual (B2C).
  – Large-scale Data Mining for social capital calculation through distributed social sensors, sales records,
    communications, web & social media activities, etc.
  – Inject historical leads and sales records to train machines to associate casualty of social capital and
    economic gains.
  – Optimize collective social & human capitals for marketing strategies and team forming.

▪ What are the related assets in Smarter Commerce solutions stack?
  – Unica Leads, NetInsight, Detect, CustomerInsight, and PredictiveInsight
  – CoreMetrics Continuous Optimization Platform

▪ What remains to be done?
  – Expand existing IBM Production System and Commercial Service Asset: SmallBlue / IBM Atlas for large-
    scale social network analysis, economic analysis, privacy-preserving data collection, marketing and sales
    expertise search and recommendation, etc, for Unica and Coremetrics platform.
  – After successful completion, sharing these Analytics with other platforms based on business needs.
An example of utilizing micro- and macro social capital

- Who among IBMers are the closest to McKinsey? What is the shortest path for me to reach McKinsey through my colleagues? Who should join the team for McKinsey to send this kind of marketing message?
- How strong is IBM, in terms of relationship strength, to other companies?
Flow of B2B Marketing using Social Network Analysis and Optimization

1) Social Network Mining
   - Among customer companies
     - From web, news articles, stocks, etc.
   - Between employee and customers
     - From intranet and sales force data
   - Among employee
     - Already available in SmallBlue

2) Quantifying Social Capital
   - Calculate the economic value of each person’s social network.
   - Analyze each person’s influence on customers or colleagues.

3) Define Constraints for Matchmaking
   - B2B (customer & customer), employee & customer
   - Condition lists for desired team should be determined based on survey for the sales force.

4) Optimize Matches based on Constraints
   - Obtain optimal or near optimal matches using Constraint Programming.
Main Steps

1. Extract a set of solution candidates (team forming) by social network analysis (SmallBlue).
2. Specify solutions that satisfy constraints by optimization from a set of solutions.

Example: Waterproof camera manufacturer

- In the past, IBM succeeded to consult company A’s to sell their waterproof camera to Best Buy.
  - Which one was the best selling model?
  - What kind of strategy lead the project to be succeeded?
- Now we have a scuba diving school B who wants to buy waterproof camera for their classes as e-marketing customer.
  It is a chance to match company A and diving school B.
- Here, this problem is defined as follows:
  1. Find social relationships between company A and diving school B.
     - Situation of relationship between A and B would be cleared by social network mining and analysis on news articles, stock markets, blogs and so on.
  2. Find following experts from IBM using SmallBlue’s social mining techniques.
     - 1 person who knows company A, 1 person who knows diving school B and 1 person who knows how to process campaign in gym, tennis school, etc.
  3. Specify appropriate people as a sales team using Tonkawa’s optimization techniques.
     - The team members are determined from a set of experts extracted by SmallBlue based on constraints like each person’s schedule, skills, expected level of their contribution, relationship among team, etc.

If we could know about available data, more scenarios would be produced.
Major technical steps for SNA for Marketing and Sales

- Make innovate Sales & Marketing software prototype (especially for B2B)

- Technical Approach
  - Foundation:
    - Privacy-Preserving Large Scale Data Mining
    - Large-Scale Network Analysis
    - Large-Scale Graph Management, Storage, Index and Retrieval
    - Large-Scale Optimization
    - Quantifiable economic and financial analysis for sales & marketing optimization strategy on graphs and networks
  - Applications:
    - Quantifying Social Capital of Customer Companies and their people:
      Finding social networks inside and outside companies by extending SmallBlue mining technology
    - Converting Social Capital into Economic Gain for B2B Marketing & Sales:
      Which employee has the shortest ‘social path’ to reach a customer company, or a specific person in a customer company? Who is the right person to send match message to customer?
      Finding matches between customer companies or between customer and employee team based on their social capitals and constraints using optimization technology

- Major Research Challenge:
  - First prototype system to quantize Social Capital, and utilize it for B2B Marketing & Sales
  - Significant amount of new system design, social & economic analytics and optimization techniques.
What keywords should I put in the search box to get the information I really want?
Multi-partite Network Analytics

Term Suggestion and Query Expansion

Document-based → Log-based → Ontology-based → Multi-partite network analytics

- Influenced by test collection characteristics
- Query log, failure for rare queries
- Click log, biased in favor of top ranks
- Not publicly available
- WordNet
- Limited semantic relatedness
- Difficult to update
- Wikipedia
- Simple concept links only
- Network community-based
- Extracting human factor
- Incorporate expertise

© 2013 Columbia University
Document-based

- Influenced by test collection characteristics
- No consideration of key terms that are highly semantically related but do not frequently co-occur.

Multi-partite Network Analytics

Term Suggestion and Query Expansion

Document-based → Log-based → Ontology-based → Multi-partite network analytics

- Influenced by test collection characteristics
- Query log, failure for rare queries
- Click log, biased in favor of top ranks
- Not publicly available
- WordNet
- Limited semantic relatedness
- Difficult to update
- Wikipedia
- Simple concept links only
- Difficult to update
- Network community-based
- Extracting human factor
- Incorporate expertise

© 2013 Columbia University
Log-based

- Cluster queries with similar clicked URLs
- Identifying the mapping between queries and clicked URLs

Multi-partite Network Analytics

Term Suggestion and Query Expansion

Document-based → Log-based → Ontology-based → Multi-partite network analytics

Influenced by test collection characteristics
Query log, failure for rare queries
Click log, biased in favor of top ranks
Not publicly available

WordNet
Limited semantic relatedness
Difficult to update

Wikipedia
Simple concept links only

Network community-based
Extracting human factor
Incorporate expertise

© 2013 Columbia University
WordNet as Ontology

- Manually constructed system based on individual words benefit will be limited
- System is not easily updated

Solar power

From Wikipedia, the free encyclopedia

This article is about generation of electricity using solar energy. For other uses of solar energy, see Solar energy.

Solar power is the conversion of sunlight into electricity, either directly using photovoltaics (PV), or indirectly using concentrated solar power (CSP). Concentrated solar power systems use lenses or mirrors and tracking systems to focus a large area of sunlight into a small beam. Photovoltaics convert light into electric current using the photoelectric effect.[1]

Commercial concentrated solar power plants were first developed in the 1980s. The 354 MW SEGS CSP installation is the largest solar power plant in the world, located in the Mojave Desert of California. Other large CSP plants include the Solnova Solar Power Station (150 MW) and the Andasol solar power station (150 MW), both in Spain. The 214 MW Charanka Solar Park in India, is the world's largest photovoltaic plant.

Contents
[hide]

1 Applications
2 Concentrating solar power
3 Photovoltaics

The PS10 concentrates sunlight from a field of heliostats onto a central tower.
Wikipedia as Ontology

- Wikipedia is a web-based free encyclopedia that anyone can edit.
- The English Wikipedia edition
  - 2.4 million articles
  - 1 billion words.
- Wikipedia relies on the power of collective intelligence
  - by peer-reviewed approaches rather than the authority of individual.
- high quality,
- almost noise free.
Previous Approaches

- Merely as an online dictionary and utilize it only as a structured knowledge database
- Using associated hyperlinks

Multi-partite Network Analytics

Term Suggestion and Query Expansion

Document-based

Log-based

Ontology-based

Multi-partite network analytics

Influenced by test collection characteristics

Query log, failure for rare queries

Click log, biased in favor of top ranks

Not publicly available

WordNet

Limited semantic relatedness

Difficult to update

Wikipedia

Simple concept links only

Network community-based

Extracting human factor

Incorporate expertise

© 2013 Columbia University
Multi-partite Network Analytics

Term Suggestion and Query Expansion

Document-based

Log-based

Ontology-based

Multi-partite network analytics

Influenced by test collection characteristics

Query log, failure for rare queries

WordNet

Wikipedia

Click log, biased in favor of top ranks

Limited semantic relatedness

Simple concept links only

Not publicly available

Difficult to update

Crawling is resource-intensive

Human factor modeling

Semantic relatedness difficult to evaluate

Our Challenge
Solar power

From Wikipedia, the free encyclopedia

Solar power is the conversion of sunlight into electricity, either directly using photovoltaics (PV), or indirectly using concentrated solar power (CSP). Concentrated solar power systems use lenses or mirrors and tracking systems to focus a large area of sunlight into a small beam. Photovoltaics convert light into electric current using the photoelectric effect.[1]

Commercial concentrated solar power plants were first developed in the 1980s. The 354 MW SEGS CSP installation is the largest solar power plant in the world, located in the Mojave Desert of California. Other large CSP plants include the S Forsmark Solar Power Station (150 MW) and the Andasol solar power station (150 MW), both in Spain. The 214 MW Charenah Solar Park in India, is the world’s largest photovoltaic plant.
Query

- Data Sampling
  - Semantic Relatedness Weighting
    - Contributor Expertise Analysis
      - Optimization
- Ontology
  - Relative Importance Ranking
    - Visualization Interface
    - Evaluation Interface
C: contributors  T: Terms  L: Categories

Layer by layer
Query

Data Sampling

Semantic Relatedness Weighting

Contributor Expertise Analysis

Optimization

Ontology

Relative Importance Ranking

Visualization Interface

Evaluation Interface

Semantic Relatedness Weighting
\[
\text{SRW} \left(t_3, t_4\right) = P\left(c_{ctb}^1\right) \cdot P\left(t_3 \mid c_{ctb}^1\right) \cdot P\left(t_4 \mid c_{ctb}^1\right) + P\left(c_{ctb}^2\right) \cdot P\left(t_3 \mid c_{ctb}^2\right) \cdot P\left(t_4 \mid c_{ctb}^2\right)
\]
Contributor to categories → Expertise → Expertise inference → Term to categories

Contributor to contributor → Expertise inference → Term to Term

Contributor Expertise factor → $C_{ctb}^{-1}$
High Semantic Relatedness Term Suggestion from Our System

"Solar Power" as keyword
Word-completion Term Suggestion

Google

- solar power
- solar power system
- solar power industries
- solar power in hong kong
- solar power bank
- solar power 2012
- solar power international 2012
- solar power hong kong
- solar power panel
- solar power in china
- solar power generation

Solar energy - Wikipedia, the free encyclopedia


www.solarpanelrebate.com/au/ - 頁庫存檔 - 翻譯這個網頁
SolarGen is a leading installer of residential solar power systems and solar panels to homes throughout Australia. Call 1300 676 527 to get your free ...
## Experiment I

<table>
<thead>
<tr>
<th></th>
<th>P@1</th>
<th>P@5</th>
<th>S@5</th>
<th>S@20</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple link</td>
<td>0.3736</td>
<td>0.3039</td>
<td>0.6017</td>
<td>0.6231</td>
<td>0.4023</td>
</tr>
<tr>
<td>+Contributor</td>
<td>0.6151</td>
<td>0.3917</td>
<td>0.8031</td>
<td>0.8116</td>
<td>0.4125</td>
</tr>
<tr>
<td>+Expertise</td>
<td>0.6693</td>
<td>0.4412</td>
<td>0.8297</td>
<td>0.9620</td>
<td>0.5919</td>
</tr>
</tbody>
</table>

Performance Comparison for Different Relationship Levels. Using BibSonomy Dataset
## Experiment II – Accuracy on different categories

<table>
<thead>
<tr>
<th>Category</th>
<th>Wordnet</th>
<th>Bag of words</th>
<th>Our algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literature</td>
<td>62.0% ± 5%</td>
<td>62.7% ± 4%</td>
<td>76.8% ± 6%</td>
</tr>
<tr>
<td>Natural science</td>
<td>60.7% ± 4%</td>
<td>65.6% ± 6%</td>
<td>73.3% ± 3%</td>
</tr>
<tr>
<td>Sociology</td>
<td>72.1% ± 5%</td>
<td>62.9% ± 5%</td>
<td>72.5% ± 7%</td>
</tr>
<tr>
<td>Business</td>
<td>60.4% ± 6%</td>
<td>58.5% ± 8%</td>
<td>67.1% ± 7%</td>
</tr>
<tr>
<td>Law</td>
<td>52.2% ± 9%</td>
<td>50.4% ± 8%</td>
<td>66.3% ± 6%</td>
</tr>
<tr>
<td>Engineering</td>
<td>54.0% ± 6%</td>
<td>68.3% ± 5%</td>
<td>66.2% ± 4%</td>
</tr>
<tr>
<td>Electrical &amp; Computer Eng.</td>
<td>77.0% ± 4%</td>
<td>68.0% ± 3%</td>
<td>82.3% ± 3%</td>
</tr>
<tr>
<td>Life Science</td>
<td>73.1% ± 6%</td>
<td>70.9% ± 6%</td>
<td>81.4% ± 7%</td>
</tr>
<tr>
<td>Agriculture</td>
<td>72.6% ± 5%</td>
<td>65.1% ± 6%</td>
<td>72.3% ± 5%</td>
</tr>
<tr>
<td>Medical</td>
<td>63.0% ± 8%</td>
<td>65.6% ± 7%</td>
<td>61.6% ± 8%</td>
</tr>
</tbody>
</table>

ODP-based precision evaluation results increase 12.5% in average
## Precision Comparison With Paraphrase Detection System

<table>
<thead>
<tr>
<th></th>
<th>Synonyms</th>
<th>Hyponymy</th>
<th>Antonyms</th>
<th>Paraphrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhao et al.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.7444</td>
</tr>
<tr>
<td>Our approach</td>
<td>0.2197</td>
<td>0.3665</td>
<td>0.2313</td>
<td>-</td>
</tr>
</tbody>
</table>

82% of the suggested terms are reported as related, *i.e.*, synonyms (22%), hyponyms (37%) or antonyms (23%)
References


