Incorporating Structured World Knowledge into Unstructured Documents via Heterogeneous Information Networks

Yangqiu Song
Outline

• Text Analytics: Motivation
  – Two Challenges
    • Representation
    • Labels

• Text Categorization via HIN
  – HIN construction from texts
  – From HIN similarity to clustering and classification
  – World knowledge indirect supervision

• Conclusions and future work
Text Categorization: Two Challenges

- Impacts many applications!
  - Social network analysis, health care, machine reading ...
- Traditional approach:
  - Two challenges:
    - Representation
    - Labels
On Feb. 8, Dong Nguyen announced that he would be removing his hit game Flappy Bird from both the iOS and Android app stores, saying that the success of the game is something he never wanted. Some fans of the game took it personally, replying that they would either kill Nguyen or kill themselves if he followed through with his decision.

Frank Lantz, the director of the New York University Game Center, said that Nguyen's meltdown resembles how some actors or musicians behave. "People like that can go a little bonkers after being exposed to this kind of interest and attention," he told ABC News. "Especially when there's a healthy dose of Internet trolls."

7 February 2014 is going to be a great day in the history of Russia with the upcoming XXII Winter Olympics 2014 in Sochi. As the climate in Russia is subtropical, hence you would love to watch ice-capped mountains from the beautiful beaches of Sochi. 2014 Winter Olympics would be an ultimate event for you to share your joys, emotions and the winning moments of your favourite sports champions. If you are really an obsessive fan of Winter Olympics games then you should definitely book your ticket to confirm your presence in winter Olympics 2014 which are going to be held in the provincial town, Sochi. Sochi Organizing committee (SOOC) would be responsible for the organization of this great international multi-sport event from 7 to 23 February 2014.
Context: Topic Models and Word Embeddings

- Topic Modeling (Blei et al., 2003)

Context: Topic Models and Word Embeddings

- Word embedding
  - Word2vec (Mikolov et al., 13)
  - Glove (Pennington et al., 14)
  - Matrix factorization (Deerwester’90; Levy et al., 15)
  - ...

https://www.tensorflow.org/versions/r0.7/tutorials/word2vec/index.html
What’s Missing?

• The semantics of entities and their relations

• What can context cover?

• What cannot?
  – Higher order relations

``New York'' vs. ``New York Times''
``George Washington'' vs. ``Washington''
Outline

• Text Analytics: Motivation
  – Two Challenges
    • Representation
    • Labels

• Text Categorization via HIN
  – HIN construction from texts
  – From HIN similarity to clustering and classification
  – World knowledge indirect supervision

• Conclusions and future work
Acquire Labeled Data

- **Expert Annotation**: Costly
- **Crowdsourcing**: Simple tasks, Low quality, Still costly
- **Semi-supervised / transfer learning**: Domain dependent

Fast changing domains

Many diverse domains

Only big companies can hire a lot of experts
Our Solution

• **World Knowledge** enabled learning
  – **Millions** of entities and concepts
  – **Billions** of relationships

• Grounding texts to knowledge bases
Classification without Supervision

- Label names carry a lot of information
  - We can use world knowledge as features
  - Classify document to English labels
  - 179 languages with Wikipedia

- July 15 08:30–09:55:
  - Machine Learning19: Classification2

---

M. Chang, L. Ratinov, D. Roth, V. Srikumar: Importance of Semantic Representation: Dataless Classification. AAAI’08.
This Talk: Structured World Knowledge Enabled Learning and Text Mining

Different domains

tweets, blogs, websites, medical, psychology

Structured world knowledge bases

[Document similarity in ICDM’15]
[Document clustering in KDD’15]
[Document classification in AAAI’16]
[Item recommendation, ongoing]

More general and effective machine learning/data mining

[Relation clustering in IJCAI’15]
[Similarity search in SDM’16]
[Paraphrasing in ACL’13]
[Data type refinement, ongoing]

With help of machine learning algorithms
Outline

• Motivation
  – Two Challenges
    • Representation
    • Labels

• Text Categorization via HIN
  – HIN construction from texts
  – From HIN similarity to clustering and classification
  – World knowledge indirect supervision

• Conclusions and future work
Text Categorization via HIN

• How to convert unstructured texts to HINs?

• What can we do with the HINs?
Challenges of Using World Knowledge

Knowledge specification; Disambiguation

Inference

Learning

Scalability; Domain adaptation; Open domain classes

Representation

Data vs. knowledge representation
Networked Text Analysis Framework

Text and World Knowledge Bases

World Knowledge Specification

World Knowledge Representation

Learning

Wang et al., Incorporating World Knowledge to Document Clustering via Heterogeneous Information Networks. KDD’15.
Wang et al. World knowledge as indirect supervision for document clustering. TKDD’16.
World Knowledge Specification: Unsupervised Semantic Parsing for Documents

Document: Obama is the president of the United States of America

Semantic parsing is the task of mapping a piece of natural language text to a formal meaning representation.

Logic form: \( \text{People.BarackObama} \sqsubseteq \text{PresidentofCountry.Country.USA} \)

Motivation: [Berant et al. EMNLP’13] aim to train a parser from question/answer pairs on a large knowledge-base Freebase

- Existing semantic parsing approaches, that require expert annotation
- Scales to large scale knowledge-bases, supervised by the QA pairs

No such training data for the document dataset.
Obama is the president of the United States of America
World Knowledge Specification: Unsupervised Semantic Parsing for Documents

Document: Obama is the president of the United States of America


Entities are linked to Freebase.

Unaries: Type.x or Profession.x.

Binaries: paths of length 1 or 2 in the KB graph.

Text phrases are from ReVerb on ClueWeb09 [Thomas Lin].

Composition rules: Join (between binary and unary); Intersection (between unary and unary).

Entities are linked to Freebase.
World Knowledge Specification: Unsupervised Semantic Parsing for Documents

Document: Obama is the president of the United States of America

- **Lexicon**: Mapping from phrases to knowledge base predicates. Unary: entity; Binary: relation.
- **Composition rules**: Join (between binary and unary); Intersection (between unary and unary).
- **Logic form construction**: based on lexicon and composition rules recursively.

Entities are linked to Freebase.

Unaries: **Type.x** or **Profession.x**.

Binaries: paths of length 1 or 2 in the KB graph.

Text phrases are from ReVerb on ClueWeb09 [Thomas Lin].
World Knowledge Specification: Unsupervised Semantic Parsing for Documents

Document: Obama is the president of the United States of America

- More than one candidate logic forms could be generated for each span of the input sentence, cannot rank.
- **Unsupervised way**
  - A state-of-art named entity recognition tool [L. Ratinov et al. CoNLL 2009] is used to find only maximum spanning phrase.
  - Only generate partial immediate logic form based on the maximum spanning phrase.

Entities are linked to Freebase.

Unaries: Type.x or Profession.x.

Binaries: paths of length 1 or 2 in the KB graph.

Text phrases are from ReVerb on ClueWeb09 [Thomas Lin].

NOT "America" or "United States"
John Smoltz came over to the Braves from the Tigers, but was developed by the Braves.

Anyhow, the Braves did try to send Bob Horner to Richmond once.

Look at Smoltz's pitching line: 6 hits, 2 walks, 1 ER, 7 SO and a loss.

Some of the forms are not noisy results.
World Knowledge Specification: Semantic Filtering

- Term frequency based semantic filtering (FBSF)
  - How many times a type appearing in a document

- Document frequency based semantic filtering (DFBSF)
  - How many documents a type appearing in, in a corpus

- Conceptualization based semantic filter (CBSF)
  - Clustering the same entity (with different mentions) based on their types
  - In each cluster, use the most frequent type for the mentions

Song et al., Open Domain Short Text Conceptualization: A Generative + Descriptive Modeling Approach. IJCAI’15.
Song et al., Short Text Conceptualization using a Probabilistic Knowledgebase. IJCAI’11.
Precision of Different Semantic Filtering

**FBSF**
*Frequency based semantic filter.*
*Type is decided by the counts in one document.*

**DFBSF**
*Document frequency based semantic filter.*
*Type is decided by the counts in whole document set.*

**CBSF**
*Conceptualization based semantic filter.*
*Type is decided by the context in whole document set.*

Wang et al., Incorporating World Knowledge to Document Clustering via Heterogeneous Information Networks. KDD’15.
Wang et al., World knowledge as indirect supervision for document clustering. TKDD’16.
Examples of Semantic Filtering on 20NG

John Smoltz came over to the Braves from the Tigers, but was developed by the Braves.

Anyhow, the Braves did try to send Bob Horner to Richmond once.

Look at Smoltz's pitching line: 6 hits, 2 walks, 1 ER, 7 SO and a loss.

---

Type.baseball_player ▷ proathlete_teams.Type.baseball_team
Type.tv_actor ▷ profession_specializations.Type.tv
Type.award_winner ▷ employment_company.Type.employer

Type.baseball_team ▷ roster_player.Type.baseball_player
Type.location ▷ contains.Type.location

proathlete_teams.Type.baseball_player
spouse_s.Type.person

John Smoltz: Type.baseball_player
Braves: Type.baseball_team
## Error Analysis of Semantic Filtering

<table>
<thead>
<tr>
<th>Type of error</th>
<th>Example sentence</th>
<th>Number and percentage of errors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>FBSF (805)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DFBsf (359)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CBSF (272)</td>
</tr>
<tr>
<td>Entity Recognition</td>
<td>“Einstein’s theory of relativity explained mercury’s motion.”</td>
<td>179 (22.2%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>129 (35.9%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>105 (38.6%)</td>
</tr>
<tr>
<td>Entity Disambiguation</td>
<td>“Bill said all this to make the point that Christianity is eminently.”</td>
<td>537 (66.7%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>182 (50.7%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>130 (47.8%)</td>
</tr>
<tr>
<td>Subordinate Clause</td>
<td>“Bruce S. Winters, worked at United States Technologies Research Center, bought a Ford.”</td>
<td>89 (11.1%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48 (13.4%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>37 (13.6%)</td>
</tr>
</tbody>
</table>

**Finding #1: Entity disambiguation is the major error factor.**

Entity disambiguation is a tough research problem in NLP community. The type information of relations are not sufficient to further prune out mismatching entities during semantic filtering process.

**Finding #2: CBSF performs the best.**

For example, by using context, the number of incorrect entities caused by disambiguation can be dramatically reduced.
Networked Text Analysis Framework

World Knowledge Specification

Text and World Knowledge Bases

Learning

World Knowledge Representation
World Knowledge Representation:
Heterogeneous Information Network (HIN)

HIN network-schema: network with multiple object types and/or multiple link types.
Outline

• Motivation
  – Two Challenges
    • Representation
    • Labels

• Text Categorization via HIN
  – HIN construction from texts
  – From HIN similarity to clustering and classification
  – World knowledge indirect supervision

• Conclusions and future work
Meta-path, Commuting Matrix, and PathSim

- Meta-path path defined over the network schema.
  - [Sun et al., 2011 ]

- Commuting matrix:
  - e.g., document->word binary occurrence matrix: $W$

- PathSim
  - e.g.,
  - $W^T W$: dot product

Bush portrayed himself as a compassionate conservative, implying he was more suitable than other Republicans to lead the United States.

Capturing higher-order relations

- Document → Politician (Contains)
  - Politician → Country (PresidentOf)
  - Politician → Document (Contains)

- Document → Baseball (Contains)
  - Baseball → Sports (Affiliation In)
  - Baseball → Document (Contains)

- Document → Military (Contains)
  - Military → Government (DepartmentOf)
  - Military → Document (Contains)
KnowSim

An ensemble of similarity measures defined on structured HIN.

Semantic overlap: the number of meta-paths between two documents.

\[ KS(d_i, d_j) = \frac{2 \times \sum_{m}^{M'} w_m |\{p_{i \rightarrow j} \in P_m\}|}{\sum_{m}^{M} w_m |\{p_{i \rightarrow i} \in P_m\}| + \sum_{m}^{M} w_m |\{p_{j \rightarrow j} \in P_m\}|} \]

Semantic broadness: the number of total meta-paths between themselves.

- **Intuition:** The larger number of highly weighted meta-paths between two documents, the more similar these documents are, which is further normalized by the semantic broadness.

- **KnowSim** is computed in nearly linear time.

Wang et al., KnowSim: A Document Similarity Measure on Structured Heterogeneous Information Networks. ICDM’15.
Challenges

# of meta-paths:
20NG (325) GCAT (1,682)

Number of meta-paths could be very large.

\[
KS(d_i, d_j) = \frac{2 \times \sum_{m}^{M'} w_m |\{p_{i\rightarrow j} \in P_m\}|}{\sum_{m}^{M'} w_m |\{p_{i\rightarrow i} \in P_m\}| + \sum_{m}^{M'} w_m |\{p_{j\rightarrow j} \in P_m\}|}
\]

The weight/importance of each meta-path is different when the domain is different.

#1: How should we generate the large number of meta-paths at the same time?
Previous studies only focus on single meta-path, enumeration over the network is OK. In real world, what will happen when thousands of meta-paths are needed?

#2: How should we decide the weight of each meta-path?
Previous studies treat them equally. In real world, different meta-path should contribute differently in various domains.
Meta-Path Dependent Random Walk

Intuition: Discovering compact sub-graph based on seed document nodes.

- Compute **Personalized PageRank (PPR)** around seed nodes.
- The random walk will get trapped inside the blue sub-graph.

**Algorithm outline**
- Run **PPR** (approximate connectivity to seed nodes) with teleport set = \{S\}
- Sort the nodes by the decreasing **PPR** score
- **Sweep** over the nodes and find compact sub-graph.
- Use the sub-graph instead of the whole graph to compute **# of meta-paths** between nodes.

Frobenius norm of approximation of commuting matrices on 20NG dataset

**Intuition:**
Discovering compact sub-graph based on seed document nodes.
Meta-Path Ranking

- Maximal Spanning Tree based Selection [Sahami, 1998]
  \[ \sum_{j \neq i}^{M} \cos(D_{.,j_1}, D_{.,j_2}) \frac{1}{M - 1} \]
  Select meta-paths with the largest dependencies with others

- Laplacian Score based Selection [He, 2006]
  \[ L_j = \frac{\hat{D}_{.,j}^T L \hat{D}_{.,j}}{\hat{D}_{.,j}^T \hat{D}_{.,j} \bowtie \hat{D}_{.,j}} \]
  Select a meta-path in discriminating documents from different clusters
### Experiments

#### Document datasets

<table>
<thead>
<tr>
<th>Name</th>
<th>#(Categories)</th>
<th>#(Leaf Categories)</th>
<th>#(Documents)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20Newsgroups (20NG)</td>
<td>6</td>
<td>20</td>
<td>20,000</td>
</tr>
<tr>
<td>MCAT (Markets)</td>
<td>9</td>
<td>7</td>
<td>44,033</td>
</tr>
<tr>
<td>CCAT (Corporate/Industrial)</td>
<td>31</td>
<td>26</td>
<td>47,494</td>
</tr>
<tr>
<td>ECAT (Economics)</td>
<td>23</td>
<td>18</td>
<td>19,813</td>
</tr>
</tbody>
</table>

MCAT, CCAT, ECAT are top categories in RCV1 dataset containing manually labeled newswire stories from Reuter Ltd.

#### World knowledge bases

<table>
<thead>
<tr>
<th>Name</th>
<th>#(Entity Types)</th>
<th>#(Entity Instances)</th>
<th>#(Relation Types)</th>
<th>#(Relation Instances)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freebase</td>
<td>1,500</td>
<td>40 millions</td>
<td>35,000</td>
<td>2 billions</td>
</tr>
<tr>
<td>YAGO2</td>
<td>350,000</td>
<td>10 millions</td>
<td>100</td>
<td>120 millions</td>
</tr>
</tbody>
</table>

publicly available knowledge base with entities and relations collaboratively collected by its community members.

a semantic knowledge base, derived from Wikipedia, WordNet and GeoNames.

The number is reported in [X. Dong et al. KDD’14], In our downloaded dump of Freebase, we found 79 domains, 2,232 types, and 6,635 properties.
Text Similarity Results

- Evaluation: correlation with document similarity
  - In the same category: 1
  - In different categories: 0

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Similarity Measures</th>
<th>BOW</th>
<th>BOW+TOPIC</th>
<th>BOW+TOPIC+ENTITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>20NG</td>
<td>Cosine</td>
<td>0.2400</td>
<td>0.2713</td>
<td>0.2768</td>
</tr>
<tr>
<td></td>
<td>Jaccard</td>
<td>0.2352</td>
<td>0.2632</td>
<td>0.2650</td>
</tr>
<tr>
<td></td>
<td>Dice</td>
<td>0.2400</td>
<td>0.2712</td>
<td>0.2767</td>
</tr>
<tr>
<td>GCAT</td>
<td>Cosine</td>
<td>0.3490</td>
<td>0.3639</td>
<td>0.3128</td>
</tr>
<tr>
<td></td>
<td>Jaccard</td>
<td>0.3313</td>
<td>0.3460</td>
<td>0.2991</td>
</tr>
<tr>
<td></td>
<td>Dice</td>
<td>0.3490</td>
<td>0.3638</td>
<td>0.3156</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>KnowSim+UNIFORM</th>
<th>KnowSim+MST</th>
<th>KnowSim+LAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>20NG</td>
<td>0.2860</td>
<td>0.2891</td>
<td>0.2913 (+5.2%)</td>
</tr>
<tr>
<td>GCAT</td>
<td>0.3815</td>
<td>0.3833</td>
<td>0.4086 (+12.3%)</td>
</tr>
</tbody>
</table>
Outline

• Motivation
  – Two Challenges
    • Representation
    • Labels

• Text Categorization via HIN
  – HIN construction from texts
  – From HIN similarity to clustering and classification
  – World knowledge indirect supervision

• Conclusions and future work
Spectral Clustering with KnowSim

- Non-linear clustering (Ng et al., NIPS’01)
  - Construct k-NN graph based on pair-wise similarities
  - Perform k-means over Eigen vectors of the graph Laplacian

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Similarity Measures</th>
<th>BOW</th>
<th>BOW+TOPIC</th>
<th>BOW+TOPIC+ENTITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>20NG</td>
<td>Cosine</td>
<td>0.3440</td>
<td>0.3461</td>
<td>0.4247</td>
</tr>
<tr>
<td></td>
<td>Jaccard</td>
<td>0.3547</td>
<td>0.3517</td>
<td>0.4292</td>
</tr>
<tr>
<td></td>
<td>Dice</td>
<td>0.3440</td>
<td>0.3457</td>
<td>0.4248</td>
</tr>
<tr>
<td>GCAT</td>
<td>Cosine</td>
<td>0.3932</td>
<td>0.4352</td>
<td>0.4106</td>
</tr>
<tr>
<td></td>
<td>Jaccard</td>
<td>0.3887</td>
<td>0.4292</td>
<td>0.4159</td>
</tr>
<tr>
<td></td>
<td>Dice</td>
<td>0.3932</td>
<td>0.4355</td>
<td>0.4112</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Datasets</th>
<th>KnowSim+UNIFORM</th>
<th>KnowSim+MST</th>
<th>KnowSim+LAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>20NG</td>
<td>0.4304</td>
<td>0.4304</td>
<td>0.4461 (+3.9%)</td>
</tr>
<tr>
<td>GCAT</td>
<td>0.4463</td>
<td>0.4653</td>
<td>0.4736 (+8.8%)</td>
</tr>
</tbody>
</table>

Wang et al., KnowSim: A Document Similarity Measure on Structured Heterogeneous Information Networks. ICDM’15.
SVM with Indefinite HIN-Kernel

- SVM needs a positive semi-definite (PSD) kernel matrix
- KnowSim matrix is non-PSD
- Feed the non-PSD KnowSim kernel matrix to SVM [Luss and d’Aspremont 2008’]
  - Learn a proxy of non-PSD KnowSim matrix
  - Simultaneously learn a SVM classifier.

Objective function:

$$\min_{\kappa, \alpha} \max_{y, \lambda} 1^T \alpha - \frac{1}{2} \alpha^T Y^T K Y \alpha + \rho \| K - K_0 \|_F^2$$

subject to:

$$y = 0, 0 \leq C, \quad K \geq 0$$
## Classification Results

### Average accuracy

<table>
<thead>
<tr>
<th>Model</th>
<th>Discrete</th>
<th>Embedding</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SVM</td>
<td>HIN</td>
<td></td>
</tr>
<tr>
<td>Settings</td>
<td>BOW</td>
<td>BOW+ENTITY</td>
<td>Word2vec</td>
</tr>
<tr>
<td>20NG-SIM</td>
<td>90.81%</td>
<td>91.11%</td>
<td>91.67%</td>
</tr>
<tr>
<td>20NG-DIF</td>
<td>96.66%</td>
<td>96.90%</td>
<td>98.27%</td>
</tr>
<tr>
<td>GCAG-SIM</td>
<td>94.15%</td>
<td>94.29</td>
<td>96.81%</td>
</tr>
<tr>
<td>GCAT-DIF</td>
<td>88.98%</td>
<td>90.18%</td>
<td>90.64%</td>
</tr>
</tbody>
</table>

### Average accuracy

<table>
<thead>
<tr>
<th>Model</th>
<th>SVM$^{\text{HIN}}$</th>
<th>SVM$^{\text{HIN}}$+KnowSim</th>
<th>IndefSVM$^{\text{HIN}}$+KnowSim</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Settings</td>
<td>DWD</td>
<td>DWD+other MetaPaths</td>
<td>DWD</td>
<td>DWD+other MetaPaths</td>
</tr>
<tr>
<td>20NG-SIM</td>
<td>91.60%</td>
<td>92.32%</td>
<td>92.68%</td>
<td>92.65%</td>
</tr>
<tr>
<td>20NG-DIF</td>
<td>97.20%</td>
<td>97.83%</td>
<td>98.01%</td>
<td>98.13%</td>
</tr>
<tr>
<td>GCAG-SIM</td>
<td>94.82%</td>
<td>95.29%</td>
<td>96.04%</td>
<td>95.63%</td>
</tr>
<tr>
<td>GCAT-DIF</td>
<td>91.19%</td>
<td>90.70%</td>
<td>91.88%</td>
<td>91.63%</td>
</tr>
</tbody>
</table>

Outline

• Motivation
  – Two Challenges
    • Representation
    • Labels

• Text Categorization via HIN
  – HIN construction from texts
  – From HIN similarity to clustering and classification
  – World knowledge indirect supervision

• Conclusions and future work
HIN Constrained Clustering Modeling

Wang et al., Incorporating World Knowledge to Document Clustering via Heterogeneous Information Networks. KDD’15.
Wang et al. World knowledge as indirect supervision for document clustering. TKDD’16.
Use the top level named entity types as the entity types in HIN.
- have a relatively dense graph.
• Use the top level named entity types as the entity types in HIN.
  – have a relatively dense graph.
• Use named entity sub-types and attributes in HIN clustering model.
  – Useful to identify the topics or clusters of the documents.
HIN Constrained Clustering Modeling

Extend the framework of information-theoretic co-clustering (ITCC) [I. S. Dhillon et al. KDD’03] and constrained ITCC [Y. Song et al. TKDE’13].

- Use the top level named entity types as the entity types in HIN.
  - have a relatively dense graph.
- Use named entity sub-types and attributes in HIN clustering model.
  - Useful to identify the topics or clusters of the documents.

Song et al. Constrained Co-clustering with Unsupervised Constraints for Text Analysis. TKDE, 2013
HIN Constrained Clustering Modeling

For documents and words, factorize

\[ q(d_m, w_i) = p(\hat{d}_k, \hat{w}_k) p(d_m | \hat{d}_k) p(w_i | \hat{w}_k) \]

Minimizing KL means approximating \( q \) should be similar to original \( p \).

\[ J_{CHINC} = D_{KL}(p(D, W) || q(D, W)) \]
\[ + \sum_{t=1}^{T} D_{KL}(p(D, E^t) || q(D, E^t)) \]
\[ + \sum_{t=1}^{T} \sum_{s=1}^{T} D_{KL}(p(E^t, E^s) || q(E^t, E^s)) \]
\[ + \sum_{t=1}^{T} \sum_{e_i^t = 1} V_t \sum_{e_i^t \in M_{e_i^t}} V_M \left( e_i^t, e_i^t \in M_{e_i^t} \right) \]
\[ + \sum_{t=1}^{T} \sum_{e_i^t = 1} V_t \sum_{e_i^t \in C_{e_i^t}} V_C \left( e_i^t, e_i^t \in C_{e_i^t} \right) \]

**Minimizing KL means approximation** \( q \) **should be similar to** original \( p \).
Clustering Algorithm

Algorithm: Alternating Optimization

Input: HIN defined on documents D, words W, entities \( E^t, t = 1, ..., T \), Set maxIter and max\( \delta \).

while iter < maxIter and \( \delta > \) max\( \delta \) do

D Label Update: minimize \( J_{CHINC} \) w.r.t. \( L_d \).

D Model Update: update \( q(d_m, w_i) \) and \( q(d_m, e_i^t) \).

for \( t = 1, ..., T \) do

\( E^t \) Label Update: minimize \( J_{CHINC} \) w.r.t. \( L_{et} \).

\( E^t \) Model Update: update \( q(d_m, e_i^t) \) and \( q(e_j^s, e_i^t) \).

end for

D Label Update: minimize \( J_{CHINC} \) w.r.t. \( L_d \).

D Model Update: update \( q(d_m, w_i) \) and \( q(d_m, e_i^t) \).

W Label Update: minimize \( J_{CHINC} \) w.r.t. \( L_w \).

W Model Update: update \( q(d_m, w_i) \).

Compute cost change \( \delta \).

end while

Knowledge indirect supervision: sub-types or attributes cannot directly affect the document labels. Constraints affect entity labels, entity labels will be transferred to affect the document labels.

The effect of different world knowledge:

- Freebase specifies more entities than YAGO2 does.

Freebase:
- specifies more entities than YAGO2 does.

World Knowledge Sources:
- Freebase
- YAGO2

Clustering NMI:
- Kmeans(BOW)
- Kmeans(BOW+YG)
- Kmeans(BOW+FB)
- ITCC(BOW)
- ITCC(BOW+YG)
- ITCC(BOW+FB)
- CITCC(BOW+ground truth)
- HINC(YG)
- HINC(FB)
- CHINC(YG)
- CHINC(FB)

Wang et al., Incorporating World Knowledge to Document Clustering via Heterogeneous Information Networks. KDD’15.
Wang et al. World knowledge as indirect supervision for document clustering. TKDD’16.
Parameter Study

Finding #1: certain values of the number of entity clusters leading to the best clustering performance.

Finding #2: larger number of iterations, the clustering improves more, and become stable. Because it comes to convergence.

Finding #3: adding more constraints leading to better performance. Then become stable. The entity sub-type information is transferred to the document side.
Other Research

• Relation search

Wang et al. RelSim: Relation Similarity Search in Schema-Rich Heterogeneous Information Networks. SDM’16.
Future Work

World knowledge bases

Different domains

tweets, blogs, websites, medical, psychology

With help of machine learning algorithms

Which domain needs to consider more structured information?

What if there is no domain knowledge in the world knowledge base?

More general and effective machine learning/data mining

[Document similarity in ICDM’15]
[Document clustering in KDD’15]
[Document classification in AAAI’16]
[Item recommendation, ongoing]

Knowledge
Networked learning

Deep learning

[Relation clustering in IJCAI’15]
[Similarity search in SDM’16]
[Paraphrasing in ACL’13]
[Data type refinement, ongoing]
## Conclusion

<table>
<thead>
<tr>
<th><strong>Problem</strong></th>
<th>Text Representation and Annotation Efforts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Framework</strong></td>
<td>World knowledge specification and representation; Text as HIN based learning and modeling</td>
</tr>
<tr>
<td><strong>System</strong></td>
<td>We are working on making analyzing text as network open source [Data and Code]</td>
</tr>
</tbody>
</table>

Thank You! 😊
Dataset

• 4 sub-datasets are constructed

<table>
<thead>
<tr>
<th>Sub-datasets</th>
<th>#(Document)</th>
<th>#(word)</th>
<th>#(Entity)</th>
<th>#(Total)</th>
<th>#(Types)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20NG-SIM</td>
<td>3000</td>
<td>22686</td>
<td>5549</td>
<td>31235</td>
<td>1514</td>
</tr>
<tr>
<td>20NG-DIF</td>
<td>3000</td>
<td>25910</td>
<td>6344</td>
<td>35254</td>
<td>1601</td>
</tr>
<tr>
<td>GCAG-SIM</td>
<td>3596</td>
<td>22577</td>
<td>8118</td>
<td>34227</td>
<td>1678</td>
</tr>
<tr>
<td>GCAT-DIF</td>
<td>2700</td>
<td>33345</td>
<td>12707</td>
<td>48752</td>
<td>1523</td>
</tr>
</tbody>
</table>

Each sub-datasets consists of three similar or distinct topics.

More entities in GCAT