Incorporating World Knowledge to Document Clustering via Heterogeneous Information Networks

KDD’15 Sydney, Australia

Chenguang Wang, Yangqiu Song, Ahmed El-Kishky, Dan Roth, Ming Zhang, Jiawei Han
Text Categorization

• A classical machine learning problem that impacts many applications!
  • Social network analysis, health care, machine reading ...
• Traditional approach:
Text Categorization

- A classical machine learning problem that impacts many applications!
  - Social network analysis, health care, machine reading ...
- Traditional approach:
  - Label data
  - Train a classifier
  - Make prediction
Text Categorization

• A classical machine learning problem that impacts many applications!
  • Social network analysis, health care, machine reading …
• Traditional approach:
  - Label data
  - Train a classifier
  - Make prediction
Challenges in Big Data Era:
How to Make Machine Learning Still Effective?
Challenges in Big Data Era: How to Make Machine Learning Still Effective?

Data Volume
Challenges in Big Data Era: How to Make Machine Learning Still Effective?

Data formal+clear
Challenges in Big Data Era: How to Make Machine Learning Still Effective?

Data Volume

Data formal+clear e.g., articles

The beginning of time would have been a point of infinite density and infinite curvature of space-time. All the known laws of physics would break down at such a point. One might suppose that there were new laws that held at singularities, but it would be very difficult even to formulate such laws at such badly behaved points, and we would have no guide from observations as to what those laws might be. However, what the singularity theorems really indicate is that the gravitational field becomes so strong that quantum gravitational effects become important. Classical theory is no longer a good description of the universe. So one has to use a quantum theory of gravity to discuss the very early stages of the universe. As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere, including at the beginning of time. It is not necessary to postulate new laws for singularities, because there need not be any singularities in the quantum theory.

We don't yet have a complete and consistent theory that combines quantum mechanics and gravity. However, we are fairly certain of some features that such a unified theory should have. One is that it should incorporate Feynman's proposal to formulate quantum theory in terms of a sum over histories. In this approach, a particle does not have just a single history, as it would in a classical theory. Instead, it is supposed to follow every possible path in
Challenges in Big Data Era: How to Make Machine Learning Still Effective?

Data Volume

As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere.

Data formal+clear e.g., articles
Challenges in Big Data Era: How to Make Machine Learning Still Effective?

Data Volume

Data

- formal+clear: e.g., articles
- informal+noisy

As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere.
Challenges in Big Data Era: How to Make Machine Learning Still Effective?

Data Volume

Data
- **formal+clear** e.g., articles
- **informal+noisy** e.g., Tweets, Queries

As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere.

KDD 2015
Challenges in Big Data Era: How to Make Machine Learning Still Effective?

Data Volume

Data

formal + clear  e.g., articles

informal + noisy  e.g., Tweets, Queries

Label

coarse-grained

As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere.

KDD 2015
Challenges in Big Data Era: How to Make Machine Learning Still Effective?

Data Volume

- **Data**
  - formal+clear: e.g., articles
  - informal+noisy: e.g., Tweets, Queries

- **Label**
  - coarse-grained

---

As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere.
Challenges in Big Data Era: How to Make Machine Learning Still Effective?

Data Volume

- **Data**
  - formal+clear: e.g., articles
  - informal+noisy: e.g., Tweets, Queries

- **Label**
  - coarse-grained
  - fine-grained

As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere.

KDD 2015

Insurance
Challenges in Big Data Era: How to Make Machine Learning Still Effective?

Data Volume

Data
- formal + clear: e.g., articles
- informal + noisy: e.g., Tweets, Queries

Label
- coarse-grained
- fine-grained

As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere.

Insurance
- Health insurance
- Auto insurance
- Vision insurance
- Dental insurance
- Disability insurance
Challenges in Big Data Era: How to Make Machine Learning Still Effective?

Data Volume

**Data**
- **formal+clear** e.g., articles
- **informal+noisy** e.g., Tweets, Queries

**Label**
- **coarse-grained**
- **fine-grained**

As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere.
Challenges in Big Data Era: How to Make Machine Learning Still Effective?

Data Volume

Data
- formal+clear e.g., articles
- informal+noisy e.g., Tweets, Queries

Label
- coarse-grained
- fine-grained

Insurance
- Health insurance
- Auto insurance
- Vision insurance
- Dental insurance
- Disability insurance

As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere.
Challenges in Big Data Era:
How to Make Machine Learning Still Effective?

Data Volume

Data
formal+clear  e.g., articles

informal+noisy  e.g., Tweets, Queries

As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere.

Label
coarse-grained

fine-grained

Health insurance
Auto insurance
...

KDD 2015

Vision insurance
Dental insurance
Disability insurance
Challenges in Big Data Era: How to Make Machine Learning Still Effective?

Data Volume

Data: formal+clear e.g., articles

As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere.

Label: coarse-grained

Machine Learning Ability

Label coarse-grained:

- Nokia
- Sony
- Dell
- Intel
- Apple

Machine Learning Ability:

Label fine-grained:

- IBM
- HP
- Microsoft
- Oracle

Machine Learning Ability:

Data Volume:

- Twitter
- Queries

Label coarse-grained:

- Insurance
  - Health insurance
  - Vision insurance
  - Dental insurance
  - Disability insurance
  - Auto insurance

Label fine-grained:

- Car
- Home
- Travel

Label coarse-grained:

- Google
- Facebook
- Twitter
- YouTube

Label fine-grained:

- Search
- Social media
- Video sharing
Challenges in Big Data Era: How to Make Machine Learning Still Effective?

Data Volume

Data: formal+clear e.g., articles

As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere

Label: coarse-grained

Nokia Sony Dell Intel Apple

Label: fine-grained

IBM HP Microsoft Oracle

More applications!

Machine Learning Ability
Acquire Labeled Data
Acquire Labeled Data

Expert Annotation

Costly
Acquire Labeled Data

Expert Annotation
Costly

Only big companies can hire a lot of experts
Acquire Labeled Data

Expert Annotation
- Costly

Crowdsourcing
- Simple tasks
- Low quality
- Still costly

Only big companies can hire a lot of experts
Acquire Labeled Data

Expert Annotation
- Costly

Crowdsourcing
- Simple tasks
- Low quality
- Still costly

Only big companies can hire a lot of experts
Acquire Labeled Data

- Expert Annotation
  - Costly

- Crowdsourcing
  - Simple tasks
  - Low quality
  - Still costly

- Semi-supervised / transfer learning
  - Not generalizable

Only big companies can hire a lot of experts
Acquire Labeled Data

Expert Annotation
- Costly

Crowdsourcing
- Simple tasks
- Low quality
- Still costly

Semi-supervised / transfer learning
- Not generalizable
- Fast changing domains

Only big companies can hire a lot of experts
Acquire Labeled Data

Expert Annotation
- Costly

Crowdsourcing
- Simple tasks
- Low quality
- Still costly

Semi-supervised /transfer learning
- Not generalizable
- Fast changing domains

Only big companies can hire a lot of experts

Many diverse domains
Knowledge Enabled Learning:
use knowledge as indirect supervision
Knowledge Enabled Learning: use knowledge as indirect supervision

World knowledge bases
Knowledge Enabled Learning: use knowledge as indirect supervision

World knowledge bases

- Wikipedia
- DBpedia
- Freebase
- Yago
- The Knowledge Graph
Knowledge Enabled Learning:
use knowledge as indirect supervision

World knowledge bases | Indirect label

- Wikipedia
- DBpedia
- YAGO
- Wikidata
- Freebase Labs
Knowledge Enabled Learning: use knowledge as indirect supervision

World knowledge bases  Indirect label

Data
Knowledge Enabled Learning: use knowledge as indirect supervision

World knowledge bases → Indirect label → Machine Learning

Data
Knowledge Enabled Learning: use knowledge as indirect supervision

World knowledge bases → Indirect label

Generalized Machine Learning

Data

Knowledge Graph

Wikipedia, DBpedia, Freebase, Yago

The Free Encyclopedia

Select knowledge

Google search

The beginning of time would have been a point of infinite laws of science would break down at such a point. One
singularities, but it would have no guide from above, indicate is that the general classical theory is not long
discuss the very early era of science to hold on to singularities, because the

We don't yet have a complete sense of what Feynman's proposal to do not have just a single hot
Knowledge Enabled Learning: use knowledge as indirect supervision

World knowledge bases ➔ Indirect label ➔ Generalized Machine Learning

Data

World knowledge bases:
- DBpedia
- Freebase Labs
- Wikipedia
- Yago

Indirect label:
- The Knowledge Graph

Generalized Machine Learning:
- select knowledge
Knowledge Enabled Learning: use knowledge as indirect supervision

- World knowledge bases
- Indirect label
- Data
- Generalized Machine Learning
- More applications
Knowledge Enabled Learning: use knowledge as indirect supervision

- World knowledge bases
- Indirect label
- More applications
- More domains
- Generalized Machine Learning
- Data

Examples:
- DBpedia
- Freebase
- Wikipedia
- Yago
- The Knowledge Graph
Example: Knowledge Enabled Text Clustering
Example: Knowledge Enabled Text Clustering


Bush portrayed himself as a compassionate conservative, implying he was more suitable than other Republicans to go to lead the United States.
Example: Knowledge Enabled Text Clustering


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

Are the two documents belong to the same cluster?
Example: Knowledge Enabled Text Clustering


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

Are the two documents belong to the same cluster?

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

Are the two documents belong to the same cluster?

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

Are the two documents belong to the same cluster? “Politics”
Example: Knowledge Enabled Text Clustering


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

Are the two documents belong to the same cluster? “Politics”

- Links and types carry a lot of information!
- But traditional approaches are not using them
Clustering of 20 Newsgroups Documents

- **Kmeans(BOW+FB)**
  Typical one-dimensional clustering Algorithm with BOW and Freebase entities as flat features.

- **ITCC(BOW+FB)**
  Information-theoretic co-clustering [I. S. Dhillon KDD’03] with BOW and Freebase entities as flat features.

- **CITCC(BOW+ground truth)**

- **HINC(FB)**
  Our method without constraints and with knowledge specified from Freebase.

- **CHINC(FB)**
  Our method with knowledge specified from Freebase.

Clustering NMI

<table>
<thead>
<tr>
<th>Method</th>
<th>NMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kmeans(BOW+FB)</td>
<td>0.447</td>
</tr>
<tr>
<td>ITCC(BOW+FB)</td>
<td>0.525</td>
</tr>
<tr>
<td>CITCC(BOW+ground truth)</td>
<td>0.569</td>
</tr>
<tr>
<td>HINC(FB)</td>
<td>0.571</td>
</tr>
<tr>
<td>CHINC(FB)</td>
<td>0.631</td>
</tr>
</tbody>
</table>
Clustering of 20 Newsgroups Documents

Clustering with knowledge: 0.631

Clustering NMI

- Kmeans(BOW+FB)
  Typical one-dimensional clustering Algorithm with BOW and Freebase entities as flat features.

- ITCC(BOW+FB)
  Information-theoretic co-clustering [I. S. Dhillon KDD'03] with BOW and Freebase entities as flat features.

- CITCC(BOW+ground truth)

- HINC(FB)
  Our method without constraints and with knowledge specified from Freebase.

- CHINC(FB)
  Our method with knowledge specified from Freebase.
Clustering of 20 Newsgroups Documents

Clustering with knowledge: 0.631

Clustering NMI

- Kmeans(BOW+FB)
  Typical one-dimensional clustering Algorithm with BOW and Freebase entities as flat features.

- ITCC(BOW+FB)
  Information-theoretic co-clustering [I. S. Dhillon KDD’03] with BOW and Freebase entities as flat features.

- CITCC(BOW+ground truth)

- HINC(FB)
  Our method without constraints and with knowledge specified from Freebase.

- CHINC(FB)
  Our method with knowledge specified from Freebase.

250K ground-truth labels as constraints.
Knowledge Enabled Learning

Challenges

Future
Knowledge Enabled Learning

Challenges

Future

World Knowledge

Data
Knowledge Enabled Learning

Challenges

General purpose problem

World Knowledge

Domain specific problem

Data

Future
Knowledge Enabled Learning

Challenges

- General purpose problem

Knowledge representation

- World Knowledge

Domain specific problem

Data representation

- Data

Future
Knowledge Enabled Learning

**Challenges**
- General purpose problem
  - Knowledge representation
    - World Knowledge
  - Large scale inference
- Domain specific problem
  - Domain representation
  - Data
  - Small scale inference

**Future**
Knowledge Enabled Learning

Challenges

Future
Knowledge Enabled Learning

Challenges

Future

More general and effective machine learning
Knowledge Enabled Learning

**Challenges**

**Future**

More applications
- e.g., clustering, classification, recommendation

More general and effective machine learning
Knowledge Enabled Learning

Challenges

Future

More applications
e.g., clustering, classification, recommendation

More general and effective machine learning

More domains
e.g., tweets, blogs, websites, medical, psychology
Knowledge Enabled Learning

Challenges

Next generation of machine learning

Future

More general and effective machine learning

More applications

e.g., clustering, classification, recommendation

More domains

e.g., tweets, blogs, websites, medical, psychology

Machine learning algorithms

World knowledge bases
Knowledge Enabled Learning

Challenges

Next generation of machine learning

Machine learning algorithms + World knowledge bases

Big data enabled machine learning

Future

More applications e.g., clustering, classification, recommendation

More general and effective machine learning

More domains e.g., tweets, blogs, websites, medical, psychology

Data
Text Clustering with World Knowledge

World knowledge bases

Documents

World Knowledge Specification

Specified World Knowledge Representation

Estimate Label

Specified World Knowledge Driven Modelling
Text Clustering with World Knowledge

World knowledge bases

Documents

World Knowledge Specification

Specified World Knowledge Representation

Specified World Knowledge Driven Modelling

Estimate Label
Unsupervised Semantic Parsing for Documents
Unsupervised Semantic Parsing for Documents

Document: Obama is the president of the United States of America
Unsupervised Semantic Parsing for Documents

DocumentObama 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.
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: \( \text{People.BarackObama} \sqsubset \text{PresidentofCountry.Country.USA} \)
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.


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

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

Logic form: \texttt{People.BarackObama} \sqcap \texttt{PresidentofCountry.Country.USA}

- Motivation: [J. 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.
Unsupervised Semantic Parsing for Documents

\[ \text{People.BarackObama} \sqcap \text{PresidentofCountry.Country.USA} \]

- lexicon
  - Obama

- lexicon
  - president
  - United States of America
Unsupervised Semantic Parsing for Documents

Unaries: Type.x or Profession.x.

Entities are linked to Freebase.


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


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

Entities are linked to Freebase.

President of United States of America

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


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.

Obama

President of Country

Country

United States of America

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

- One than one candidate logic forms could generate 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.

- People.BarackObama
- PresidentofCountry.Country.USA

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].

Lexicon

• 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.
Semantic Filtering

• Conceptualization based semantic filter (CBSF).

Assumption: correct semantic meaning can best fit the context.
Different entities can be used to disambiguate each other.
Semantic Filtering

• Conceptualization based semantic filter (CBSF).

Assumption: correct semantic meaning can best fit the context.
Different entities can be used to disambiguate each other.
Semantic Filtering

• Conceptualization based semantic filter (CBSF).

Assumption: correct semantic meaning can best fit the context.
Different entities can be used to disambiguate each other.

apple    adobe
Semantic Filtering

• Conceptualization based semantic filter (CBSF).

Assumption: correct semantic meaning can best fit the context. Different entities can be used to disambiguate each other.

apple
 software company, brand, fruit

adobe
 brand, software company
Semantic Filtering

• Conceptualization based semantic filter (CBSF).

Assumption: correct semantic meaning can best fit the context. Different entities can be used to disambiguate each other.

apple
software company, brand, fruit

adobe
brand, software company
Semantic Filtering

• Conceptualization based semantic filter (CBSF).

Assumption: correct semantic meaning can best fit the context.
Different entities can be used to disambiguate each other.

\[
P(\text{type} \mid \text{related entities})
\]

Song et al. Short text conceptualization using a probabilistic knowledgebase. IJCAI’11.
Semantic Filtering

- Conceptualization based semantic filter (CBSF).

Assumption: correct semantic meaning can best fit the context.
Different entities can be used to disambiguate each other.

Song et al. Short text conceptualization using a probabilistic knowledgebase. IJCAI’11.
Semantic Filtering

• Conceptualization based semantic filter (CBSF).

Assumption: correct semantic meaning can best fit the context. Different entities can be used to disambiguate each other.

Assumption: correct semantic meaning can best fit the context. Different entities can be used to disambiguate each other.

\[
P(\text{type} | \text{related entities})
\]

A cluster of entities of type features

Song et al. Short text conceptualization using a probabilistic knowledgebase. IJCAI’11.
Examples of Semantic Filtering on 20NG

Logic Forms

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

proathleteTeams.Type.baseball_player
spouse_s.Type.person

Filtered Semantics

John Smoltz:Type.baseball_player
Braves:Type.baseball_team
Text Clustering with World Knowledge

World Knowledge Specification

Specified World Knowledge Representation

Specified World Knowledge Driven Modelling

World knowledge bases
Documents

Estimate Label
Document-based Heterogeneous Information Network (HIN)

HIN: Network with multiple object types and/or multiple link types.
Document-based Heterogeneous Information Network (HIN)

HIN: Network with multiple object types and/or multiple link types.

Y. Sun et al. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. PVLDB’11.
A good way to model real world data!

Document-based Heterogeneous Information Network (HIN)

HIN: Network with multiple object types and/or multiple link types.

Y. Sun et al. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. PVLDB’11.
A good way to model real world data!

Network schema: High-level description of a network.

Y. Sun et al. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. PVLDB’11.
Document-based Heterogeneous Information Network (HIN)

HIN: Network with multiple object types and/or multiple link types.

A good way to model real world data!

Network schema: High-level description of a network.

Y. Sun et al. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. PVLDB’11.
**Document-based Heterogeneous Information Network (HIN)**

**HIN:** Network with multiple object types and/or multiple link types.

A good way to model real world data!

Network schema: High-level description of a network.

Y. Sun et al. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. PVLDB’11.
Document-based Heterogeneous Information Network (HIN)

HIN: Network with multiple object types and/or multiple link types.

A good way to model real world data!

Network schema: High-level description of a network.

Y. Sun et al. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. PVLDB’11.
Document-based Heterogeneous Information Network (HIN)

HIN: Network with multiple object types and/or multiple link types.

A good way to model real world data!

Network schema: High-level description of a network.

Y. Sun et al. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. PVLDB’11.
Text Clustering with World Knowledge

World knowledge bases

Documents

World Knowledge Specification

Specified World Knowledge Representation

Specified World Knowledge Driven Modelling

Estimate Label
Constrained Clustering Modeling

![Diagram of a Hierarchical Information Network (HIN) with partition for Doc Cluster 1 and Doc Cluster 2]

- Document
  - Word
  - Named Entity Type 1
  - Named Entity Type 2
  - Named Entity Type 3
  - Named Entity Type T

HIN partition

Doc Cluster 1

Doc Cluster 2
Constrained Clustering Modeling
Constrained Clustering Modeling

- Use the top level named entity types as the entity types in HIN.
  - have a relatively dense graph.
Constrained Clustering Modeling

- 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.
Constrained Clustering Modeling

- 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.
Constrained Clustering Modeling

- 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.
Constrained Clustering Modeling

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

\[ q(d_m, w_i) = p(d_{kd}, \hat{w}_k) p(d_m|d_{kd}) p(w_i|\hat{w}_k) \]
Constrained Clustering Modeling

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

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

- Joint probability
- \( p(d_m, w_i) \) approximation
- Cluster indicators
- Cluster indices
Constrained Clustering Modeling

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

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

Joint probability
\(p(d_m, w_i)\) approximation
Cluster indicators
Cluster indices

\[
J_{CHINC} = D_{KL}(p(D, W)||q(D, W)) + \sum_{t=1}^{T} D_{KL}(p(D, E^t)||q(D, E^t))
+ \sum_{t=1}^{T} \sum_{s=1}^{T} D_{KL}(p(E^t, E^s)||q(E^t, E^s))
+ \sum_{t=1}^{T} \sum_{e_{i_1}^t=1}^{V_t} \sum_{e_{i_2}^t \in M_{e_{i_1}^t}} V_M(e_{i_1}^t, e_{i_2}^t \in M_{e_{i_1}^t})
+ \sum_{t=1}^{T} \sum_{e_{i_1}^t=1}^{V_t} \sum_{e_{i_2}^t \in C_{e_{i_1}^t}} V_C(e_{i_1}^t, e_{i_2}^t \in C_{e_{i_1}^t})
\]
Motivation: The framework of information-theoretic co-clustering (ITCC) [I. S. Dhillon et al. KDD’03] and constrained ITCC [Y. Song et al. TKDE’13].
Constrained Clustering Modeling

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

$$q(d_m, w_i) = p(\hat{d}_{kd}, \hat{w}_{kw}) p(d_m | \hat{d}_{kd}) p(w_i | \hat{w}_{kw})$$

Minimize

$$J_{CHINC} = D_{KL}(p(D, W) || q(D, W)) + \sum_{t=1}^{T} D_{KL}(p(D, E^t) || q(D, E^t))$$

$$+ \sum_{t=1}^{T} \sum_{s=1}^{T} D_{KL}(p(E^t, E^s) || q(E^t, E^s))$$

$$+ \sum_{t=1}^{T} \sum_{e_{i_1}^t = 1}^{V_t} \sum_{e_{i_2}^t \in M_{e_{i_1}^t}} V_M (e_{i_1}^t, e_{i_2}^t \in M_{e_{i_1}^t})$$

$$+ \sum_{t=1}^{T} \sum_{e_{i_1}^t = 1}^{V_t} \sum_{e_{i_2}^t \in C_{e_{i_1}^t}} V_C (e_{i_1}^t, e_{i_2}^t \in C_{e_{i_1}^t})$$
Constrained Clustering Modeling

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

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

Minimize

\[
J_{CHINC} = D_{KL}(p(D, W)||q(D, W)) + \sum_{t=1}^T D_{KL}(p(D, E^t)||q(D, E^t))
\]

\[
+ \sum_{t=1}^T \sum_{s=1}^T D_{KL}(p(E^t, E^s)||q(E^t, E^s))
\]

\[
+ \sum_{t=1}^T \sum_{v_t=e_{i_1}^t=1}^{V_t} \sum_{e_{i_2}^t \in M_{e_{i_1}^t}} V_M(e_{i_1}^t, e_{i_2}^t \in M_{e_{i_1}^t})
\]

\[
+ \sum_{t=1}^T \sum_{v_t=e_{i_1}^t=1}^{V_t} \sum_{e_{i_2}^t \in C_{e_{i_1}^t}} V_C(e_{i_1}^t, e_{i_2}^t \in C_{e_{i_1}^t})
\]
Constrained Clustering Modeling

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

\[ q(d_m, w_i) = p(\tilde{d}_{kd}, \tilde{w}_{kw}) p(d_m | \tilde{d}_{kd}) p(w_i | \tilde{w}_{kw}) \]

Minimize \( J_{CHINC} = D_{KL}(p(D, W) || q(D, W)) + \sum_{t=1}^{T} D_{KL}(p(D, E^t) || q(D, E^t)) + \sum_{t=1}^{T} \sum_{s=1}^{T} D_{KL}(p(E^t, E^s) || q(E^t, E^s)) + \sum_{t=1}^{T} \sum_{e_{i_1}^t = 1}^{V_t} \sum_{e_{i_2}^t \in M_{e_{i_1}^t}} V_M(e_{i_1}^t, e_{i_2}^t \in M_{e_{i_1}^t}) + \sum_{t=1}^{T} \sum_{e_{i_1}^t = 1}^{V_t} \sum_{e_{i_2}^t \in C_{e_{i_1}^t}} V_C(e_{i_1}^t, e_{i_2}^t \in C_{e_{i_1}^t}) \)
Constrained Clustering Modeling

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

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

Minimize

\[ J_{\text{CHINC}} = D_{KL}(p(D,W) || q(D,W)) + \sum_{t=1}^{T} p(D,E^t) || q(D,E^t) \]

\[ + \sum_{t=1}^{T} \sum_{s=1}^{T} D_{KL}(p(E^t,E^s) || q(E^t,E^s)) \]

\[ + \sum_{t=1}^{T} \sum_{e_{i_1}^t=1}^{V_t} \sum_{e_{i_2}^t \in M_{e_{i_1}^t}} V_M(e_{i_1}^t, e_{i_2}^t \in M_{e_{i_1}^t}) \]

\[ + \sum_{t=1}^{T} \sum_{e_{i_1}^t=1}^{V_t} \sum_{e_{i_2}^t \in C_{e_{i_1}^t}} V_C(e_{i_1}^t, e_{i_2}^t \in C_{e_{i_1}^t}) \]

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

Motivation: The framework of information-theoretic co-clustering (ITCC) [I. S. Dhillon et al., KDD'03] and constrained ITCC [Y. Song et al., TKDE'13].

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

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

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

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

Minimize

\[ J_{CHINC} = D_{KL}(p(D, W)||q(D, W)) + \sum_{t=1}^{T} D_{KL}(p(D, E^t)||q(D, E^t)) \]

\[ + \sum_{t=1}^{T} \sum_{s=1}^{T} D_{KL}(p(E^t, E^s)||q(E^t, E^s)) \]

\[ + \sum_{t=1}^{T} \sum_{i_1=1}^{V_t} \sum_{i_2=1}^{e_{i_2} \in M_{e_{i_1}}} V_M(e_{i_1}, e_{i_2} \in M_{e_{i_1}}) \]

\[ w_M D_{KL}(p(D|e_{i_1}^t)||p(D|e_{i_2}^t)) \cdot l_{i_1} \neq l_{i_2} \]

\[ + \sum_{t=1}^{T} \sum_{i_1=1}^{V_t} \sum_{i_2=1}^{e_{i_2} \in C_{e_{i_1}}} V_C(e_{i_1}, e_{i_2} \in C_{e_{i_1}}) \]

\[ w_C (D_{max} - D_{KL}(p(D|e_{i_1}^t)||p(D|e_{i_2}^t))) \cdot l_{i_1} \neq l_{i_2} \]

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

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

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

Joint probability \( p(d_m, w_i) \) approximation

Cluster indicators

Cluster indices

Minimize

\[ J_{CHINC} = D_{KL}(p(D, W) || q(D, W)) \]

\[ + \sum_{t=1}^{T} D_{KL}(p(D, E^t) || q(D, E^t)) \]

\[ + \sum_{t=1}^{T} \sum_{s=1}^{T} D_{KL}(p(E^t, E^s) || q(E^t, E^s)) \]

\[ + \sum_{t=1}^{T} \sum_{e_i^t=1}^{V_t} \sum_{e_i^t \in M_{e_i^t}} V_M(e_i^t, e_i^t \in M_{e_i^t}) \]

\[ + \sum_{t=1}^{T} \sum_{e_i^t=1}^{V_t} \sum_{e_i^t \in C_{e_i^t}} V_C(e_i^t, e_i^t \in C_{e_i^t}) \]

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

Knowledge indirect supervision: fine-grained named entity sub-types or the attributes

Must-link: if two labels are not equal, consider how dissimilar they are

Cannot-link: if two labels are equal, consider how similar they are
Constrained Clustering Modeling

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

\[
q(d_m, w_i) = p(d_k | w_k)p(d_m | k)p(w_i | w_k)
\]

\[
\sum_{t=1}^{T} \sum_{t=1}^{T} D_{KL}(p(D, W) || q(D, W))
\]

Minimizing KL means approximation q should be similar to original p.

Knowledge indirect supervision: fine-grained named entity sub-types or the attributes

Minimize

Cannot-link: if two labels are equal, consider how similar they are

Must-link: if two labels are not equal, consider how dissimilar they are
Constrained Clustering Modeling

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

Cluster indicators
Cluster indices

Joint probability
\( p(d_m, w_i) \) approximation

Globally optimizing the latent labels and the approximating function is intractable

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 \( J_{\text{CHINC}} \) w.r.t. \( L_d \).
  D Model Update: update \( q(d_m, w_i) \) and \( q(d_m, e^t_I) \).
  for \( t = 1, ..., T \) do
    E_t Label Update: minimize \( J_{\text{CHINC}} \) w.r.t. \( L_{e_t} \).
    E_t Model Update: update \( q(d_m, e_t^I) \) and \( q(e_t^I, e_t^I) \).
  end for

  D Label Update: minimize \( J_{\text{CHINC}} \) w.r.t. \( L_d \).
  D Model Update: update \( q(d_m, w_i) \) and \( q(d_m, e^t_I) \).

  W Label Update: minimize \( J_{\text{CHINC}} \) w.r.t. \( L_w \).
  W Model Update: update \( q(d_m, w_i) \).

Compute cost change \( \delta \).
end while

Minimize

\[
J_{\text{CHINC}} = D_{KL}(p(D, W) || q(D, W)) + \sum_{t=1}^{T} D_{KL}(p(D, E^t) || q(D, E^t)) \]

+ \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_{t_1} = 1}^{V_t} \sum_{e_{t_2} \in M_{e_{t_1}^t}} V_M(e_{t_1}, e_{t_2}^t) \]

+ \sum_{t=1}^{T} \sum_{e_{t_1} = 1}^{V_t} \sum_{e_{t_2} \in C_{e_{t_1}^t}} V_C(e_{t_1}, e_{t_2}^t)

Must-link: if two labels are not equal, consider how dissimilar they are

Cannot-link: if two labels are equal, consider how similar they are

\[
q(d_m, w_i) = p(d_k^d, \hat{w}_k^w) p(d_m | d_k^d) p(w_i | \hat{w}_k^w)
\]
Constrained Clustering Modeling

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

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

Globally optimizing the latent labels and the approximating function is intractable

Algorithm: Alternating Optimization
Input: HIN defined on documents D, words W, entities \(E^t, t = 1, ..., T\), Set maxiter and maxδ.
while iter < maxiter and δ > maxδ do
    D Label Update: minimize \(J_{CHINC}\) w.r.t. \(L_d\).
    D Model Update: update \(q(d_m, w_i)\) and \(q(d_m, e^t_i)\).
    for \(t = 1, ..., T\) do
        \(E^t\) Label Update: minimize \(J_{CHINC}\) w.r.t. \(L_{e^t}\).
        \(E^t\) Model Update: update \(q(d_m, e^t_i)\) and \(q(e^t_i, e^t_f)\).
    end for
    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
Constrained Clustering Modeling

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

Minimize

\[ J_{CHINC} = D_{KL}(p(D, W) || q(D, W)) + \sum_{t=1}^{T} D_{KL}(p(D, E^t) || q(D, E^t)) + \sum_{t=1}^{T} \sum_{s=1}^{T} D_{KL}(p(E^t, E^s) || q(E^t, E^s)) + \sum_{t=1}^{T} V_M(e^t_{i_1}, e^t_{i_2} \in M_{e^t_{i_1}}) + \sum_{t=1}^{T} V_C(e^t_{i_1}, e^t_{i_2} \in C_{e^t_{i_1}}) \]

Global optimizing the latent labels and the approximating function is intractable

Algorith: Alternating Optimization

Input: HIN defined on documents D, words W, entities \( E^t, t = 1, \ldots, T \). Set maxiter and maxδ.

while iter < maxiter and δ > maxδ do

D Label Update: minimize \( J_{CHINC} \) w.r.t. \( L_d \).
D Model Update: update \( q(d_m, w_i) \) and \( q(d_m, e^t_i) \).

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

\( E^t \) Label Update: minimize \( J_{CHINC} \) w.r.t. \( L_{e^t} \).
\( E^t \) Model Update: update \( q(d_m, e^t_i) \) and \( q(e^t_i, e^t_j) \).
end for

D Label Update: minimize \( J_{CHINC} \) w.r.t. \( L_d \).
D Model Update: update \( q(d_m, w_i) \) and \( q(d_m, e^t_i) \).

W Label Update: minimize \( J_{CHINC} \) w.r.t. \( L_w \).
W Model Update: update \( q(d_m, w_i) \).

Compute cost change δ.
end while
Motivation: The framework of information-theoretic co-clustering (ITCC) [I. S. Dhillon et al. KDD'03] and constrained ITCC [Y. Song et al. TKDE'13].

Joint probability $p(d_m, w_i)$ approximation

Cluster indicators

Cluster indices

Minimize

Globally optimizing the latent labels and the approximating function is intractable

Algorithm: Alternating Optimization

Input: HIN defined on documents $D$, words $W$, entities $E^t$, $t = 1, ..., T$. Set maxIter and maxδ.

while iter < maxIter and δ > maxδ do

for $t = 1, ..., T$ do

$E^t$ Label Update: minimize $J_{CHINC}$ w.r.t. $L_d$.

$E^t$ Model Update: update $q(d_m, e^t)$ and $q(d_m, e^t_{i_1})$.

end for

D Label Update: minimize $J_{CHINC}$ w.r.t. $L_d$.

D Model Update: update $q(d_m, w_i)$ and $q(d_m, e^t)$.

W Label Update: minimize $J_{CHINC}$ w.r.t. $L_w$.

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

Compute cost change $δ$.

end while
Constrained Clustering Modeling

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

$$q(d_m, w_i) = p(d_{kd}, w_{kw})p(d_m | d_{kd})p(w_i | w_{kw})$$

$$J_{CHiNC} = D_{KL}(p(D, W) || q(D, W)) + \sum_{t=1}^T D_{KL}(p(D, E^t) || q(D, E^t))$$

Must-link: if two labels are not equal, consider how dissimilar they are

Cannot-link: if two labels are equal, consider how similar they are

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 $\delta > \text{max}\delta$ do

for $t = 1, ..., T$ do

$E^t$ Label Update: minimize $J_{CHiNC}$ w.r.t. $L_d^t$.

$E^t$ Model Update: update $q(d_m, e^t_i)$ and $q(d_m, e^t_i)$. 

end for

D Label Update: minimize $J_{CHiNC}$ w.r.t. $L_d$.

D Model Update: update $q(d_m, w_i)$ and $q(d_m, e^t_i)$. 

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
Constrained Clustering Modeling

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

Joint probability $p(d_m, w_i)$ approximation

Cluster indicators

Cluster indices

Minimize

$$J_{CHINC} = D_{KL}(p(D, W)||q(D, W)) + \sum_{t=1}^{T} D_{KL}(p(D, E^t)||q(D, E^t)) + \sum_{t=1}^{T} \sum_{s=1}^{T} D_{KL}(p(E^t, E^s)||q(E^t, E^s))$$

Must-link: if two labels are not equal, consider how dissimilar they are

Cannot-link: if two labels are equal, consider how similar they are

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.

Document

Word

Named Entity Type 1

Named Entity Type 2

Named Entity Type 3

Named Entity Type T

Globally optimizing the latent labels and the approximating function is intractable

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 $\delta > \text{max}\delta$ do

D Label Update: minimize $J_{CHINC}$ w.r.t. $L_d$.

D Model Update: update $q(d_m, w_i)$ and $q(d_m, e^t_i)$.

for $t = 1, ..., T$ do

$E^t$ Label Update: minimize $J_{CHINC}$ w.r.t. $L_{e^t}$.

$E^t$ Model Update: update $q(d_m, e^t_i)$ and $q(e^t_i, e^s_i)$.

end for

D Label Update: minimize $J_{CHINC}$ w.r.t. $L_d$.

D Model Update: update $q(d_m, w_i)$ and $q(d_m, e^t_i)$.

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
# 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>

## 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>

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

YAGO2 - a semantic knowledge base, derived from Wikipedia, WordNet and GeoNames.
# 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>

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

YAGO2 is a semantic knowledge base, derived from Wikipedia, WordNet and GeoNames.
## 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.
# 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>

- Freebase is a publicly available knowledge base with entities and relations collaboratively collected by its community members.
- YAGO2 is 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.
Clustering Results on 20 Newsgroups

Clustering with Freebase knowledge performs the best: 0.631
Clustering Results on 20 Newsgroups

Clustering with Freebase knowledge performs the best: 0.631

Clustering NMI
Clustering Results on 20 Newsgroups

Clustering with Freebase knowledge performs the best: 0.631
Clustering Results on 20 Newsgroups

Clustering with Freebase knowledge performs the best: 0.631
Clustering Results on 20 Newsgroups

Clustering with Freebase knowledge performs the best: 0.631
Clustering Results on 20 Newsgroups

Clustering with Freebase knowledge performs the best: 0.631

- **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)

**BOW features**
- **BOW + YAGO2 entity features**
- **BOW + Freebase entity features**

**Information-theoretic co-clustering (ITCC)** [I. S. Dhillon KDD’03]

**Constrained information-theoretic co-clustering** [Y. Song TKDE’13] with BOW + 250K ground-truth labels.

**Our method without constraints**
- Use Freebase as world knowledge.
- Use YAGO2 as world knowledge.

**Our method**
Clustering Results on 20 Newsgroups

Clustering with Freebase knowledge performs the best: 0.631

- 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)

- BOW features
- BOW + YAGO2 entity features
- BOW + Freebase entity features
- Information-theoretic co-clustering (ITCC) [I. S. Dhillon KDD'03]
- Our method without constraints
- Use Freebase as world knowledge.
- Use YAGO2 as world knowledge.
- Our method

250K ground-truth labels.
Clustering Results on 20 Newsgroups

Clustering with Freebase knowledge performs the best: 0.631

The effect of different world knowledge

- Freebase specifies more entities than YAGO2 does

Clustering with Freebase knowledge performs the best: 0.631

BOW features
BOW + YAGO2 entity features
BOW + Freebase entity features
Information-theoretic co-clustering (ITCC) [I. S. Dhillon KDD’03]
Our method without constraints
Use Freebase as world knowledge.
Use YAGO2 as world knowledge.
Our method

0.429 0.437 0.447
0.501 0.513 0.525
0.569 0.541 0.571
0.6 0.631
Parameter Study

Clustering with different numbers of entity clusters of each entity type

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

Clustering with different numbers of entity clusters of each entity type

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.
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 and more constraints leading to better performance. Then become stable. The entity sub-type information is transferred to the document side.
<table>
<thead>
<tr>
<th>Problem</th>
<th>Document clustering with world knowledge as indirect supervision.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Framework</td>
<td>World knowledge specification: unsupervised semantic parsing and conceptualization based semantic filtering.</td>
</tr>
<tr>
<td>Model</td>
<td>Constrained clustering model with the specified world knowledge represented in heterogeneous information network.</td>
</tr>
</tbody>
</table>
### Problem
Document clustering with world knowledge as indirect supervision.

### Framework
World knowledge specification: unsupervised semantic parsing and conceptualization based semantic filtering.

### Model
Constrained clustering model with the specified world knowledge represented in heterogeneous information network.

Thank You! 😊