Aspect Extraction with Automated Prior Knowledge Learning

Zhiyuan (Brett) Chen
Arjun Mukherjee
Bing Liu
Aspect Extraction

1. Extracting aspect terms
This camera takes beautiful pictures but its price is higher than $200.
This camera takes beautiful pictures but its price is higher than $200.
Aspect Extraction

1. Extracting aspect terms

2. Grouping terms into categories
Grouping

- Picture
- Photo
- Image

Aspect 1

 Aspect 2

- Price
- Cost
- Money
Aspect Extraction

Input: A review collection

Output: A set of aspects (with top aspect terms).

Aspect 1
- Price
- Cheap
- Cost
- Money
- Pricy

Aspect 2
- Battery
- Life
- Charge
- AAA
- Hour
Topic Models to Extract Aspects
(e.g., Chen et al., 2013; Kim et al., 2013; Lazaridou et al., 2013; Mukherjee and Liu, 2012; Moghaddam and Ester, 2011; Sauper et al., 2011; Lin and He, 2009; Titov and McDonald, 2008; Lu and Zhai, 2008;)

Perform both extracting and grouping

A topic is basically an aspect
Traditional Modeling Flow

Domain 1

$M$ Docs
Traditional Modeling Flow

Domain 1

$M$ Docs

LDA

$T$ Topics
Traditional Modeling Flow

Domain 1

\[ M \text{ Docs} \]

\[ \text{LDA} \]

\[ T \text{ Topics} \]

Domain 2

\[ M \text{ Docs} \]

\[ \text{LDA} \]

\[ T \text{ Topics} \]
Traditional Modeling Flow

Domain 1

\[ M \text{ Docs} \]

\[ \text{LDA} \]

\[ T \text{ Topics} \]

Domain 2

\[ M \text{ Docs} \]

\[ \text{LDA} \]

\[ T \text{ Topics} \]

Domain \( N \)

\[ M \text{ Docs} \]

\[ \text{LDA} \]

\[ T \text{ Topics} \]

\[ \ldots \]
Can we improve these topics by using them only?
Can we improve these topics by using them only?

Fully automatic

No other resources

No human intervention
Our Proposed Algorithm

Domain 1

$M$ Docs

LDA

$T$ Topics

Domain 2

$M$ Docs

LDA

$T$ Topics

Domain $N$

$M$ Docs

LDA

$T$ Topics

...
Our Proposed Algorithm

Domain 1

\( M \) Docs

LDA

\( T \) Topics

Domain 2

\( M \) Docs

LDA

\( T \) Topics

…

Domain \( N \)

\( M \) Docs

LDA

\( T \) Topics

Topic Base
Our Proposed Algorithm

Domain 1

$M$ Docs

LDA

$T$ Topics

Learn Knowledge Automatically

Knowledge Base

Domain $N$

$M$ Docs

LDA

$T$ Topics

Topic Base
Our Proposed Algorithm

Domain 1: $M$ Docs → LDA → $T$ Topics

Domain 2: $M$ Docs → LDA → $T$ Topics

... Domain $N$: $M$ Docs → LDA → $T$ Topics

Topic Base

Learn Knowledge Automatically

Knowledge Base

a) Existing Domains
Our Proposed Algorithm

Domain 1
- M Docs
  - LDA
  - T Topics

Domain 2
- M Docs
  - LDA
  - T Topics

... 

Domain N
- M Docs
  - LDA
  - T Topics

Topic Base

Learn Knowledge Automatically

Knowledge Base

a) Existing Domains

Domain 1
- M Docs
  - AKL
  - T Topics

Domain 2
- M Docs
  - AKL
  - T Topics

Domain N
- M Docs
  - AKL
  - T Topics

AKL (Automated Knowledge LDA)
Our Proposed Algorithm

Domain 1
- $M$ Docs
- LDA
- $T$ Topics

Domain 2
- $M$ Docs
- LDA
- $T$ Topics

... 

Domain $N$
- $M$ Docs
- LDA
- $T$ Topics

Topic Base

b) New Domain

Knowledge Base

Learn Knowledge Automatically
Our Proposed Algorithm

Domain 1: $M$ Docs
   - LDA
   - $T$ Topics

Domain 2: $M$ Docs
   - LDA
   - $T$ Topics

... Domain $N$:
   - LDA
   - $T$ Topics

Learn Knowledge Automatically

Knowledge Base

b) New Domain

Domain $N+1$:
   - $M$ Docs
   - AKL
   - $T$ Topics
Why don’t we merge documents from different domains and run LDA?
Run LDA on Merged Data

Number of Topics

Topic belongs to which domain

Scalability
Run LDA on Merged Data
Run LDA on Merged Data
Our Proposed Algorithm

Run LDA

Run LDA

Run LDA

Run LDA

Run LDA

Run LDA
Our Proposed Algorithm

Learn Knowledge
Our Proposed Algorithm

Run AKL
Run AKL
Run AKL
Run AKL
Run AKL
Run AKL
Run AKL
Learn Knowledge Automatically

ISSUES

Multiple Senses

Knowledge
Reliability
Learn Knowledge Automatically

ISSUES

Multiple Senses

Knowledge
Reliability
Multiple Senses

Light

\{\text{Light, Bright}\}
\{\text{Light, Luminance}\}

\{\text{Light, Weight}\}
\{\text{Light, Heavy}\}
Existing Models with Multiple Senses

Assume single sense
DF-LDA (Andrzejewski et al., 2009)

User specified multiple senses
MC-LDA (Chen et al., 2013)

Automatically distinguish senses when extracting knowledge
Learn knowledge Automatically

ISSUES

Multiple Senses

Solution

Knowledge

Reliability

Topic Clustering
Topic Clustering

A topic represents words with similar meaning (but noisy)

Group topics with similar sense into one cluster

Different senses of a word should be split into different clusters
Learn knowledge Automatically

ISSUES

Multiple Senses ➔ Topic Clustering

Solution

Knowledge

Reliability
Topic Overlapping

Every product domain has price.

Most electronic domains have battery.

Some electronic domains share screen.
Example

D1
 Battery
 Life
 Picture
 Charge

D2
 Battery
 Price
 Life
 Size

D3
 Battery
 Charge
 AAA
 Screen
Example

D1
Battery
Life
Picture
Charge

D2
Battery
Price
Life
Size

D3
Battery
Charge
AAA
Screen

Two words together at least 2 times
Two words together at least 2 times

{Battery, Life} and {Battery, Charge}
Learn knowledge Automatically

ISSUES

Multiple Senses

Knowledge Reliability

Solution

Topic Clustering

Frequent Itemset Mining

Learn knowledge Automatically
Frequent Itemset Mining (FIM)

Each topic is a transaction

Find frequent patterns satisfy minimum support thresholds

Each pattern contains 2 terms
Knowledge Representation

In the form of knowledge clusters (KC)

Each KC has a list of frequent 2-patterns

KC1: {battery, life}, {battery, charge}, {battery, hour}, {charge, hour}
AKL (Automated Knowledge LDA)

ISSUES

Incorporate Knowledge

Wrong Know.
Towards Domain
AKL Model

ISSUES

Incorporate Knowledge

Wrong Know.
Towards Domain

Solution

Add variable $c$
AKL Plate Notation

c: knowledge cluster
AKL Plate Notation

\[ c: \text{knowledge cluster} \]
c: knowledge cluster
AKL Plate Notation

c: knowledge cluster
AKL Model

ISSUES

Incorporate Knowledge

Wrong Know.
Towards Domain

Solution

Add variable $c$

GPU Model
LDA with SPU (Simple Pólya Urn Model)
LDA with SPU  (Simple Pólya Urn Model)
AKL with GPU (Generalized Pólya Urn Model)
AKL with GPU (Generalized Pólya Urn Model)

\{price, cheap\}

Topic 0
AKL Model

ISSUES
Incorporate Knowledge
Wrong Know.
Towards Domain

Solution
Add variable $c$
GPU Model
Wrong Know. Towards Domain

Wrong because of TM mistakes
\{Price, Picture\}

Wrong towards a particular domain
\{Light, Bright\}
\{Light, Weight\}
AKL Model

ISSUES

Incorporate Knowledge

Wrong Know. Towards Domain

Solution

Add variable \( c \)

GPU Model

Co-Document Frequency Ratio
Co-Doc \((w, w')\) = \frac{D(w, w') + 1}{(D(w) + D(w')) \times \frac{1}{2} + 1}
Co-Document Frequency Ratio

\[
C_{o-Doc}(w, w') = \frac{D(w, w') + 1}{(D(w) + D(w')) \times \frac{1}{2} + 1}
\]

Estimated in the current domain
Co-Document Frequency Ratio

\[ Co-Doc(w, w') = \frac{D(w, w') + 1}{(D(w) + D(w'))} \times \frac{1}{2} + 1 \]

Estimated in the current domain

\{Price, Cheap\}
\{Price, Image\}
Evaluation
Evaluation

36 product domains. Each domain:

1000 Reviews
15 Topics
Model Comparison

LDA (Blei et al., 2003)

MC-LDA (Chen et al., 2013)

GK-LDA (Chen et al., 2013)
Model Comparison

LDA (Blei et al., 2003)

MC-LDA (Chen et al., 2013)

GK-LDA (Chen et al., 2013)

Feed them with the knowledge from our algorithm
## Example Aspects

<table>
<thead>
<tr>
<th>Camera</th>
<th>Computer</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AKL</strong></td>
<td><strong>AKL</strong></td>
</tr>
<tr>
<td>battery</td>
<td>battery</td>
</tr>
<tr>
<td>life</td>
<td>hour</td>
</tr>
<tr>
<td>hour</td>
<td>hour</td>
</tr>
<tr>
<td>long</td>
<td>life</td>
</tr>
<tr>
<td>charge</td>
<td>long</td>
</tr>
<tr>
<td>extra</td>
<td>speaker</td>
</tr>
<tr>
<td>minute</td>
<td>sound</td>
</tr>
<tr>
<td>charger</td>
<td>connection</td>
</tr>
<tr>
<td>short</td>
<td>life</td>
</tr>
<tr>
<td>aa</td>
<td>hdmus</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>LDA</strong></th>
<th><strong>LDA</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>card</td>
<td>cable</td>
</tr>
<tr>
<td>memory</td>
<td>speaker</td>
</tr>
<tr>
<td>life</td>
<td>dvi</td>
</tr>
<tr>
<td>usb</td>
<td>sound</td>
</tr>
<tr>
<td>hour</td>
<td>tv</td>
</tr>
<tr>
<td>minute</td>
<td>hdmus</td>
</tr>
<tr>
<td>sd</td>
<td>tv</td>
</tr>
<tr>
<td>extra</td>
<td></td>
</tr>
<tr>
<td>device</td>
<td></td>
</tr>
</tbody>
</table>
Human Evaluation

![Chart with precision values for different devices and methods]

- **Camera**: AKL 0.9, GK-LDA 0.85, MC-LDA 0.8, LDA 0.75
- **Computer**: AKL 0.85, GK-LDA 0.8, MC-LDA 0.75, LDA 0.7
- **Headphone**: AKL 0.8, GK-LDA 0.75, MC-LDA 0.7, LDA 0.65
- **GPS**: AKL 0.8, GK-LDA 0.75, MC-LDA 0.7, LDA 0.65
Number of Topic Clusters

![Graph showing the relationship between number of clusters and topic coherence. The x-axis represents the number of clusters ranging from 20 to 70, while the y-axis represents topic coherence ranging from -1510 to -1430. The graph indicates a trend where topic coherence increases with the number of clusters up to a certain point and then decreases.]
Conclusions

To extract better aspects

Learn knowledge automatically

AKL: Leverage automated knowledge
Learn knowledge Automatically

ISSUES

Multiple Senses
Knowledge Reliability

Solution

Topic Clustering
Frequent Itemset Mining
AKL Model

ISSUES

Incorporate Knowledge

Wrong Know. Towards Domain

Solution

Add variable c

GPU Model

Co-Document Frequency Ratio
Q&A

Thank you!