Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction

Xiaojun Wan, Jianwu Yang and Jianguo Xiao

Institute of Computer Science and Technology, Peking University, Beijing, 100871, China

April 14, 2010
Introduction

Methods

Experiment Results

Conclusions

Reference
Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction

Xiaojun Wan, Jianwu Yang and Jianguo Xiao

Outline
Introduction
Methods
Experiment
Results
Conclusions
Reference

Information Overload

- Everyday huge amount of information is generated
Information Overload

- Everyday huge amount of information is generated
- Automatic analysis and extraction techniques are needed
  - Information retrieval
  - Text mining
Creating a compressed version of a given document
- Document Summarization: sentence
- Keyword Extraction: word
Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction

Xiaojun Wan, Jianwu Yang and Jianguo Xiao

Outline
Introduction
Methods
Experiment
Results
Conclusions
Reference

Document Summarization/Keyword Extraction

- Creating a compressed version of a given document
  - Document Summarization: sentence
  - Keyword Extraction: word

- Task
  - Query-relevant
  - Generic
Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction

Xiaojun Wan, Jianwu Yang and Jianguo Xiao

Outline
- Introduction
- Methods
- Experiment
- Results
- Conclusions
- Reference

Document Summarization/Keyword Extraction

- Creating a compressed version of a given document
  - Document Summarization: sentence
  - Keyword Extraction: word

- Task
  - Query-relevant
  - Generic

- Approach
  - Extraction
  - Abstraction
Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction

Outline
- Introduction
- Methods
- Experiment
- Results
- Conclusions
- Reference

Assign saliency score to sentences and then rank all the candidates
Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction

Xiaojun Wan, Jianwu Yang and Jianguo Xiao

Outline
Introduction
Methods
Experiment
Results
Conclusions
Reference

Illustrations: Document Summarization

- Assign saliency score to sentences and then rank all the candidates

- Approaches:
  - Scoring by statistical/linguistic features, maximizing marginal relevance, etc. [1, 3, 5]
Assign saliency score to words and then rank all the candidates.

text/14854 china industrial output rises in first quarter china's industrial output rose 14.1 pct in the first quarter of 1987 against the same 1986 period, the people's daily said. its overseas edition said the growth rate, which compares with a target of seven pct for the whole of 1987, was "rather high" but the base in the first quarter of 1986 was on the low side. industrial output grew 4.4 pct in the first quarter of 1986. it said china's industrial production this year has been normal but product quality and efficiency need further improvement. it gave no further details.
Assign saliency score to words and then rank all the candidates.

**Approaches:**
- Syntactic clues, phrase classification, etc. [2, 6]
Assumption.1
A sentence should be salient if it is heavily linked with other salient sentences, and a word should be salient if it is heavily linked with other salient words.

Assumption.2
A sentence should be salient if it contains many salient words, and a word should be salient if it appears in many salient sentences.
Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction

Xiaojun Wan, Jianwu Yang and Jianguo Xiao

Outline
- Introduction
- Methods
- Experiment
- Results
- Conclusions
- Reference

Intuitive Illustration

- **SS-Relationship:** homogeneous relationships between sentences
- **WW-Relationship:** homogeneous relationships between words
- **SW-Relationship:** heterogeneous relationships between sentences and words

![Diagram showing relationships between sentences (SS), words (WW), and sentences and words (SW).]
Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction

Xiaojun Wan, Jianwu Yang and Jianguo Xiao

Outline
Introduction
Methods
Experiment Results
Conclusions
Reference

Intuitive Illustration

SS-Relationship: homogeneous relationships between sentences

WW-Relationship: homogeneous relationships between words
Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction

Xiaojun Wan, Jianwu Yang and Jianguo Xiao

Outline
Introduction
Methods
Experiment
Results
Conclusions
Reference

Intuitive Illustration

- SS-Relationship: homogeneous relationships between sentences
- WW-Relationship: homogeneous relationships between words
- SW-Relationship: heterogeneous relationships between sentences and words
Graph Construction: SS-Graph

- **Edge**: content similarity based

\[
U_{ij} = \begin{cases} 
    \frac{\vec{s}_i \times \vec{s}_j}{||\vec{s}_i|| \times ||\vec{s}_j||} & \text{if } i \neq j \\
    0 & \text{otherwise}
\end{cases}
\]

where \(\vec{s}_i \times \vec{s}_j\) is the cosine similarity between sentence \(s_i\) and \(s_j\).
Graph Construction: SS-Graph

- Edge: content similarity based
  \[ U_{ij} = \begin{cases} 
  \frac{s_i \times s_j}{||s_i|| \times ||s_j||} & \text{if } i \neq j \\
  0 & \text{otherwise} 
  \end{cases} \]

  where \( s_i \times s_j \) is the cosine similarity between sentence \( s_i \) and \( s_j \).

- Direction: normalization by rows
  \[ \tilde{U}_{ij} = \begin{cases} 
  \frac{U_{ij}}{\sum_{j=1}^{m} U_{ij}} & \text{if } \sum_{j=1}^{m} U_{ij} \neq 0 \\
  0 & \text{otherwise} 
  \end{cases} \]
Graph Construction: WW-Graph

- Edge: word similarity
Graph Construction: WW-Graph

- Edge: word similarity
Graph Construction: WW-Graph

- **Edge:** word similarity
  - Corpus-based measure: mutual information [7]

\[
sim(t_i, t_j) = \log \frac{N \times p(t_i, t_j)}{p(t_i) \times p(t_j)}
\]
Graph Construction: WW-Graph

- Edge: word similarity
  - Corpus-based measure: mutual information [7]

\[
sim(t_i, t_j) = \log \frac{N \times p(t_i, t_j)}{p(t_i) \times p(t_j)}
\]

- Direction: normalization by rows
  - make the sum of each row equal to 1
Graph Construction: SW-Graph

- Bipartite graph

\[
aff(s_i, t_j) = \frac{tf_{t_j} \times isf_{t_j}}{\sum_{t \in S_i} tf_{t_j} \times isf_{t_j}}
\]

where tf and isf is the term frequency in the sentence and the inverse sentence frequency.
Graph Construction: SW-Graph

- Bipartite graph

\[
aff(s_i, t_j) = \frac{tf_{t_j} \times isf_{t_j}}{\sum_{t \in S_i} tf_{t_j} \times isf_{t_j}}
\]

where \( tf \) and \( isf \) is the term frequency in the sentence and the inverse sentence frequency.

- Edge:
  - \( \hat{W}_{ij} \) Normalize \( aff(s_i, t_j) \) by rows
  - \( \tilde{W}_{ij} \) Normalize \( aff(s_i, t_j) \) by columns
Reinforcement Algorithm

- Assumption 1: information propagation in homogeneous networks

\[ u(s_i) \propto \sum_j \tilde{U}_{ji} u(s_j) \]

\[ v(t_j) \propto \sum_i \tilde{V}_{ij} v(t_i) \]
Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction

Xiaojun Wan, Jianwu Yang and Jianguo Xiao

Outline
Introduction
Methods
Experiment
Results
Conclusions
Reference

Reinforcement Algorithm

- Assumption 1: information propagation in homogeneous networks
  \[ u(s_i) \propto \sum_j \tilde{U}_{ji} u(s_j) \]
  \[ v(t_j) \propto \sum_i \tilde{V}_{ij} v(t_i) \]

- Assumption 2: information propagation in heterogeneous network
  \[ u(s_i) \propto \sum_j \hat{W}_{ji} v(t_j) \]
  \[ v(t_j) \propto \sum_i \hat{W}_{ij} u(s_i) \]
Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction

Reinforcement Algorithm

- **Iterative updating:**

\[
\begin{align*}
u(s_i)^{(n)} &= \alpha \sum_j \tilde{U}_{ji} u(s_j)^{(n-1)} + \beta \sum_j \hat{W}_{ji} \nu(t_j)^{(n-1)} \\
\nu(t_j)^{(n)} &= \alpha \sum_i \hat{W}_{ij} u(s_i)^{(n-1)} + \beta \sum_i \tilde{W}_{ij} u(s_i)^{(n-1)}
\end{align*}
\]
Evaluation Setup
Document Summarization

- **Data set**
  - Evaluation set: DUC2002, 567 English news articles
  - Background corpus: DUC2001-2005

- **Measurement**
  - ROUGE: overlapping between the candidate summary and the reference summary
### Evaluation Results

#### Document Summarization

<table>
<thead>
<tr>
<th>System</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-W</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Approach (WN)</td>
<td>0.47100*</td>
<td>0.20424*</td>
<td>0.16336*</td>
</tr>
<tr>
<td>Our Approach (MI:k=2)</td>
<td>0.46711*</td>
<td>0.20195*</td>
<td>0.16257*</td>
</tr>
<tr>
<td>Our Approach (MI:k=5)</td>
<td>0.46803*</td>
<td>0.20259*</td>
<td>0.16310*</td>
</tr>
<tr>
<td>Our Approach (MI:k=10)</td>
<td>0.46823*</td>
<td>0.20301*</td>
<td>0.16294*</td>
</tr>
<tr>
<td>SentenceRank</td>
<td>0.45591</td>
<td>0.19201</td>
<td>0.15789</td>
</tr>
<tr>
<td>MutualRank</td>
<td>0.43743</td>
<td>0.17986</td>
<td>0.15333</td>
</tr>
</tbody>
</table>

**Table 1. Summarization Performance before Removing Redundancy (w/o MMR)**

<table>
<thead>
<tr>
<th>System</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-W</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Approach (WN)</td>
<td>0.47329*</td>
<td>0.20249*</td>
<td>0.16352*</td>
</tr>
<tr>
<td>Our Approach (MI:k=2)</td>
<td>0.47281*</td>
<td>0.20281*</td>
<td>0.16375*</td>
</tr>
<tr>
<td>Our Approach (MI:k=5)</td>
<td>0.47282*</td>
<td>0.20249*</td>
<td>0.16343*</td>
</tr>
<tr>
<td>Our Approach (MI:k=10)</td>
<td>0.47223*</td>
<td>0.20225*</td>
<td>0.16308*</td>
</tr>
<tr>
<td>SentenceRank</td>
<td>0.46261</td>
<td>0.19457</td>
<td>0.16018</td>
</tr>
<tr>
<td>MutualRank</td>
<td>0.43805</td>
<td>0.17253</td>
<td>0.15221</td>
</tr>
</tbody>
</table>

**Table 2. Summarization Performance after Removing Redundancy (w/ MMR)**
Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction

Xiaojun Wan, Jianwu Yang and Jianguo Xiao

Outline
Introduction
Methods
Experiment Results
Conclusions
Reference

Parameter Sensitivity
Document Summarization

Figure 2. ROUGE-1 vs. $\alpha$

Figure 3. ROUGE-2 vs. $\alpha$
Evaluation Setup
Keyword Extraction

- Data set
  - Evaluation set: DUC2002, 34 documents in the first 5 clusters
  - Manually labeling: 10 salient words for each document

- Measurement
  - Precision/Recall: treating as a classification problem
## Evaluation Results

### Keyword Extraction

<table>
<thead>
<tr>
<th>System</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Approach (WN)</td>
<td>0.413</td>
<td>0.504</td>
<td>0.454</td>
</tr>
<tr>
<td>Our Approach (MI:k=2)</td>
<td>0.428</td>
<td>0.485</td>
<td>0.455</td>
</tr>
<tr>
<td>Our Approach (MI:k=5)</td>
<td>0.425</td>
<td>0.491</td>
<td>0.456</td>
</tr>
<tr>
<td>Our Approach (MI:k=10)</td>
<td>0.393</td>
<td>0.455</td>
<td>0.422</td>
</tr>
<tr>
<td>WordRank (k=2)</td>
<td>0.373</td>
<td>0.412</td>
<td>0.392</td>
</tr>
<tr>
<td>WordRank (k=5)</td>
<td>0.368</td>
<td>0.422</td>
<td>0.393</td>
</tr>
<tr>
<td>WordRank (k=10)</td>
<td>0.379</td>
<td>0.407</td>
<td>0.393</td>
</tr>
<tr>
<td>MutualRank</td>
<td>0.355</td>
<td>0.397</td>
<td>0.375</td>
</tr>
</tbody>
</table>

Table 3. The Performance of Keyword Extraction
A running example

- Document ID: D062/AP891018-0301
- Labeled keywords:
  - insurance earthquake insurer damage california Francisco pay
- Extracted keywords:
  - WN: insurance earthquake insurer quake california spokesman cost million wednesday damage
  - MI(k=5): insurance insurer earthquake percent benefit california property damage estimate rate
Conclusion and Future Work

- A unified framework for fusing sentence-to-sentence, word-to-word, sentence-to-word relationships
Conclusion and Future Work

- A unified framework for fusing sentence-to-sentence, word-to-word, sentence-to-word relationships
- Other similarity measurement could be investigated to construct the relation network
Conclusion and Future Work

- A unified framework for fusing sentence-to-sentence, word-to-word, sentence-to-word relationships
- Other similarity measurement could be investigated to construct the relation network
- Apply it to multi-document summarization and keyword extraction
Thank you!
Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction

E. Hovy and C. Lin.
Automated text summarization in SUMMARIST.

B. Krulwich and C. Burkey.
Learning user information interests through extraction of semantically significant phrases.

T. Nomoto and Y. Matsumoto.
A new approach to unsupervised text summarization.

References
T. Pedersen, S. Patwardhan, and J. Michelizzi.
Wordnet:: similarity-measuring the relatedness of concepts.

D. Shen, J. Sun, H. Li, Q. Yang, and Z. Chen.
Document summarization using conditional random fields.

A. Steier and R. Belew.
Exporting phrases: A statistical analysis of topical language.
Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction

Xiaojun Wan, Jianwu Yang and Jianguo Xiao

References

P. Turney.
Learning algorithms for keyphrase extraction.