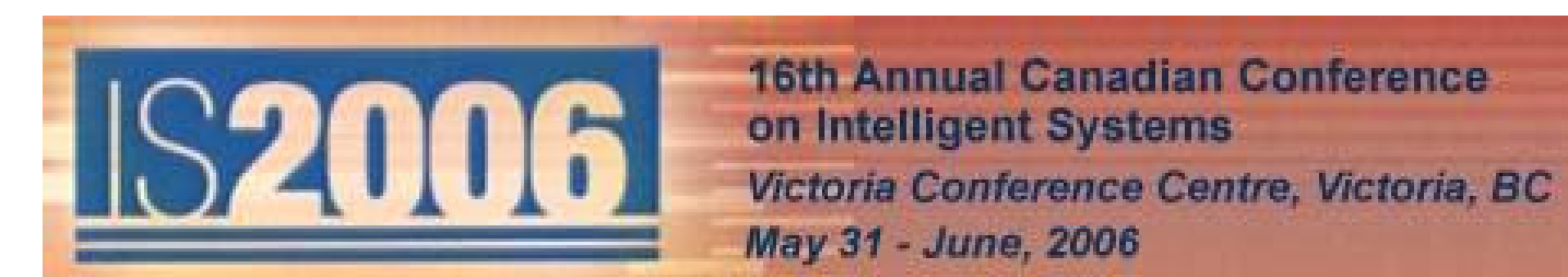


A Framework for Summarizing Multi-topic Web Sites

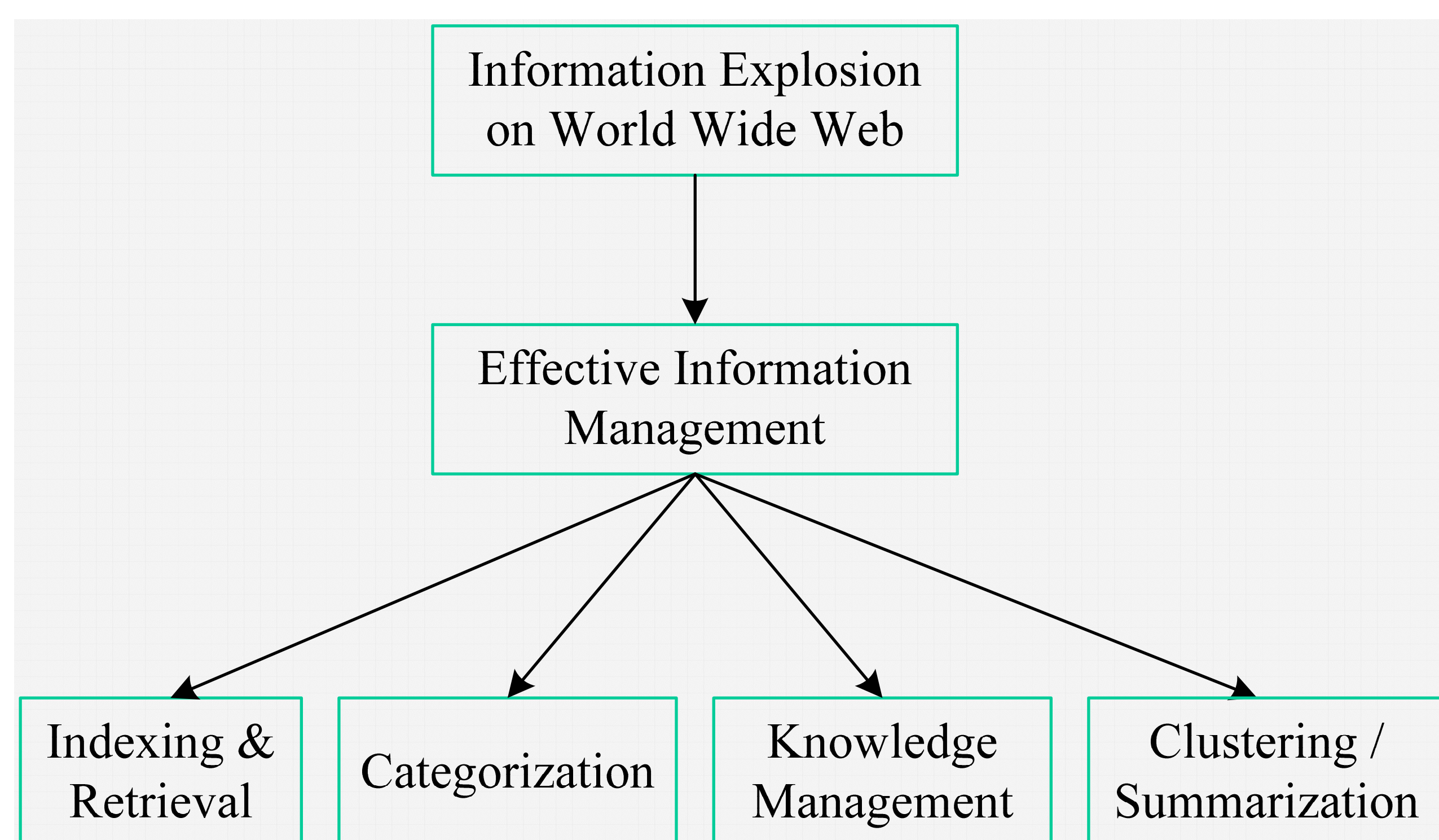
Yongzheng Zhang, Evangelos Milios, and Nur Zincir-Heywood

Faculty of Computer Science, Dalhousie University

<http://www.cs.dal.ca/~{yongzhen, eem, zincir}>



1. Web Information Management



2. Why Web Page Summarization?

- **Single Web Page Summarization**
 - Query formation and expansion
 - Web page display on handheld devices
 - Web document indexing
 - Document relevance ranking
- **Multiple Web Page Summarization**
 - Navigation and browsing of a Web site
 - Organization of search engine results
 - Organization of product reviews
 - Web directory construction (Yahoo!, DMOZ)

3. Direct Summarization of Entire Web Site [1]

- Generate a brief yet informative **summary**
- Key components:
 - **Narrative text classification:** C5.0 learning
 - **Key phrase extraction:** C-value/NC-value term likelihood
 - **Key sentence extraction:** the most essential topic
- **Problem:** unable to cover multiple topics

4. Clustering and Summarization Framework [2]

- **Clustering** of documents to find main topics:
 - **Text-based:** feature selection
 - Document Frequency (**DF**)
 - Term Variance (**TV**)
 - **Link-based:** co-citation and bibliographic coupling
 - **X-means** clustering using Vector Space Model
 - Test data sets: CMU SEI lab, MIT CSAIL, Air Canada
 - Evaluation: Precision, Recall, F1-measure
- **Summarization** of individual clusters
 - Key sentence classification: C5.0 learning
 - Part-of-speech statistics, page depth, length, distance
- Top 5 cluster summaries

5. Summary Examples

- A single long summary

Part I. Top 25 Key Phrases (show top 15)		
1. Software engineering institute	6. Product line	11. Software engineering
2. Carnegie mellon university	7. Software product	12. Software architecture
3. Capability maturity model	8. Software process	13. Model integration
4. Personal software process	9. Partner network	14. Capability maturity
5. General navigation button	10. Carnegie mellon	15. Process improvement

Part II. Top 25 Key Sentences (show top 3)
1. The Software Engineering Institute (SEI) sponsors, co-sponsors, and is otherwise involved in many events throughout the year.
2. The Software Engineering Institute offers a number of courses and training opportunities.
3. Information contained on the Software Engineering Institute Web site is published in the interest of scientific and technical information exchange.

- Top 5 short cluster summaries (show top 3)

Each has 5 key phrases & 5 key sentences (show top 3 & top 1)

Cluster I. Software Architecture		
1. Software architecture	2. software engineering	3. architecture professional
SEI architecture experts provide technical assistance and coaching in software architecture requirements, software architecture design, software architecture documentation, and architecture-centric life-cycle practices.		

Cluster II. Risk Management		
1. Risk management	2. risk management team	3. program risk
The Team Risk Management Service extends Continuous Risk Management overview to all organizations in a program, tailoring methods and tools to the joint management of program risks.		

Cluster III. People CMM		
1. People CMM	2. capability maturity model	3. capability maturity
The People Capability Maturity Model (People CMM) is a framework that helps organizations successfully address their critical people issues.		

6. Experiments and Evaluation

- **User study:** how informative two types of summaries are
 - 20 graduate students rank relatedness of summary elements to the most essential topic using a 1-to-5 scale
- **Quality measure:** acceptable percentage
 - Ratio of elements with scores 3, 4, or 5
- **ANOVA test:** top 5 cluster summaries significantly outperforms a single long summary

7. Contributions and Conclusions

- **An intelligent system** for Web site summarization
 - Coupled text and link analysis
 - Able to find and summarize essential topics
 - Effective organization and visualization of documents
- Future work: Hierarchical summarization

References:

- [1] Y. Zhang, N. Zincir-Heywood, E. Milios, Summarizing Web Sites Automatically. Canadian AI 2003, Best Paper Award!
- [2] Y. Zhang, Topic Hierarchy Construction for Web Site Summarization. SIGIR 2005 Doctoral Consortium.

Acknowledgements: NSERC, ITIS, GINIus