Multi-Dimensional Summarization In Cyber-Physical Society

Hai Zhuge

http://www.knowledgegrid.net/~h.zhuge

CGW2016 Keynote

Human Summarization

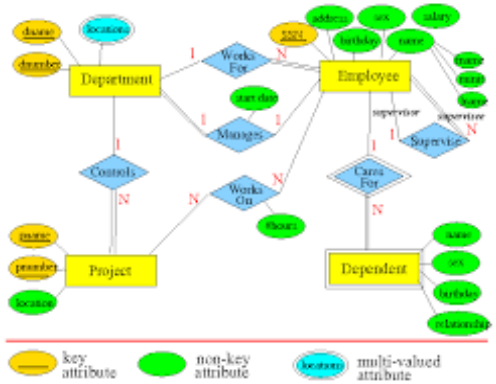
IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 6, JUNE 2009 785

Communities and Emerging Semantics in Semantic Link Network: Discovery and Learning

Hai Zhuge, *Senior Member, IEEE*

Abstract—The World Wide Web provides plentiful contents for Web-based learning, but its hyperlink-based architecture connects Web resources for browsing freely rather than for effective learning. To support effective learning, an e-learning system should be able to discover and make use of the semantic communities and the emerging semantic relations in a dynamic complex network of learning resources. Previous graph-based community discovery approaches are limited in ability to discover semantic communities. This paper first suggests the Semantic Link Network (SLN), a loosely coupled semantic data model that can semantically link resources and derive out implicit semantic links according to a set of relational reasoning rules. By studying the intrinsic relationship between semantic communities and the semantic space of SLN, approaches to discovering reasoning-constraint, rule-constraint, and classification-constraint semantic communities are proposed. Further, the approaches, principles, and strategies for discovering emerging semantics in dynamic SLNs are studied. The basic laws of the semantic link network motion are revealed for the first time. An e-learning environment incorporating the proposed approaches, principles, and strategies to support effective discovery and learning is suggested.

Contents



Problems of Text Summarization

- Too specific to apply to various forms of text
  - Words, phrases, and sentences — indicators rather than semantics
  - Interpretability is weak (no theory)
  - Knowledge needed for a good summarization is beyond text
  - Why can human use words beyond text?
- Neglect human behavior
  - Writers and readers experience with representation and summarization in lifetime.
  - Machines do not have the ability of experiencing
- Closed system
  - The process of summarization is closed
  - In cyberspace, reviews and comments are open, easily available, and valuable for improving summarization

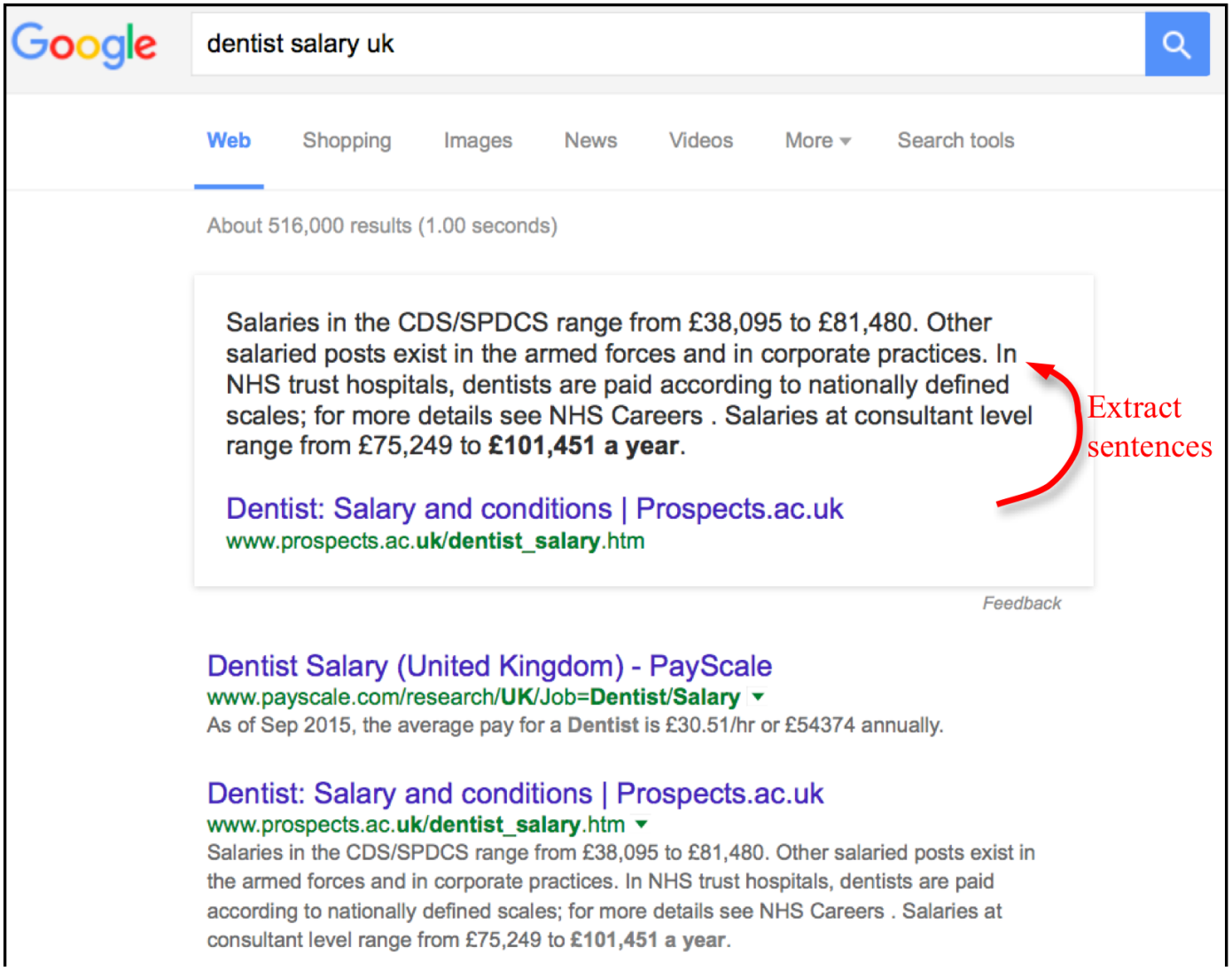
Automatic Text Summarization

- Problem
  - Input a representation and output a short representation
- General solution
  - Selection
    - Scan and select the important language components (e.g., sentences) from source text according to some measures.
  - Ordering
    - Determine the order of the selected components.
  - Composition
    - Compose the selected components to get a new text, where components should be organized for understanding easily.

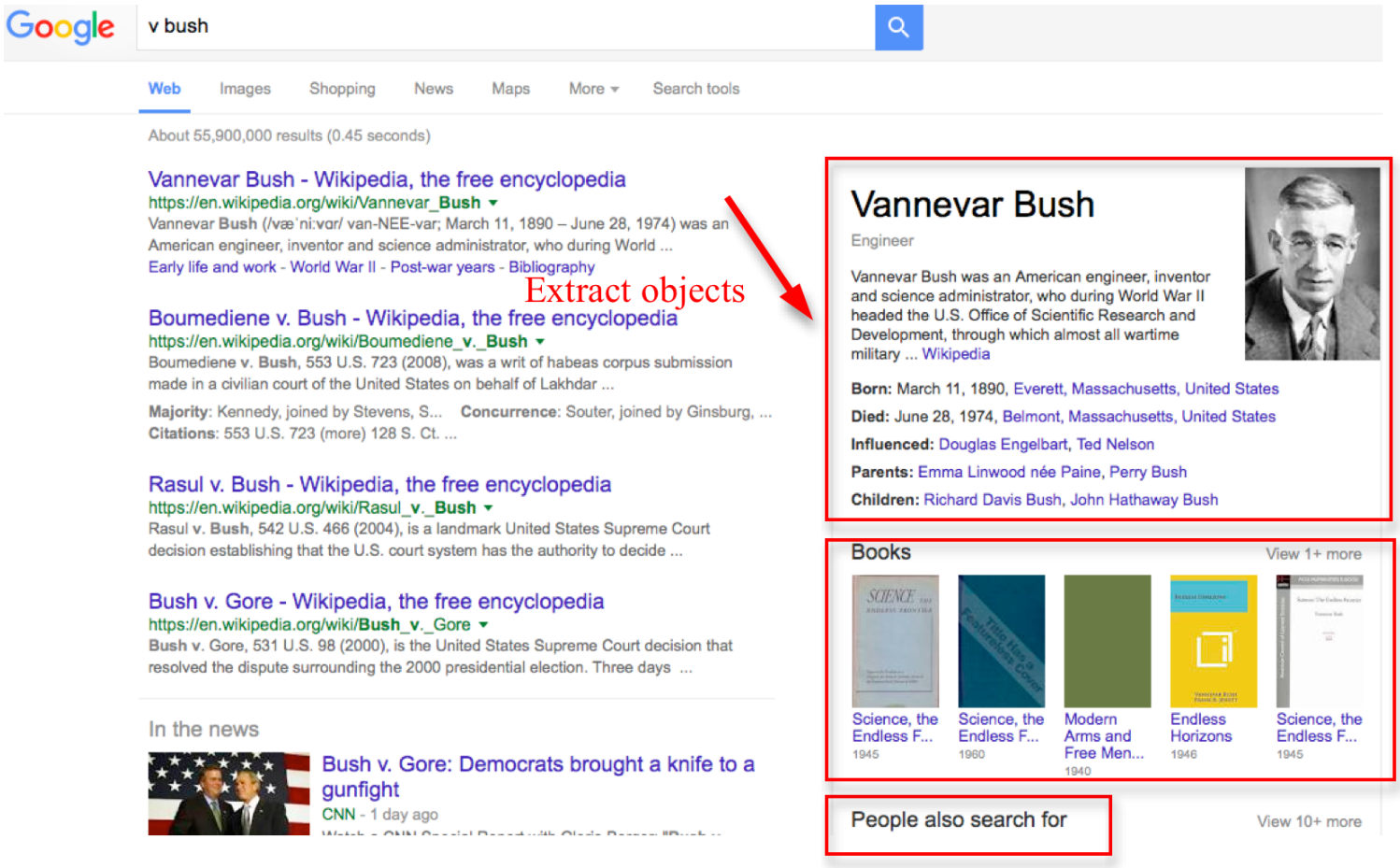
It is a pleasure to write a foreword to Hai Zhuge's new book and to welcome him to the UK as a colleague, as he takes up his chair at Aston University. Within the wider Semantic Web research project, I think Hai Zhuge is trying to do something bold and important: first, by asking how can big data, and all that that movement now entails, be made to yield up representations that give us some understanding of the content implicit in the data, but is not essentially textual. This is his Resource Space Model, a multi-dimensional category space that uniformly represents human understanding and the patterns in real and abstract objects. Secondly, he is asking how can we have a common representation for the content of different modalities: not just speech and text but images and 3D objects, the long-awaited Internet of Things. He then takes summarization as a reasonably well-understood task environment where the notion of summarization beyond the textual, the standard case so far, can be explored and ultimately tested. I think he is asking serious and fundamental questions in this work and proposing bold research solutions. This is an important issue in the history of idea development of the semantic web where many have transitioned seamlessly from the term "semantic web" to the "data web," without always thinking of what that entails. For example, the whole motivation of the semantic web was to have a web of understandable meanings or interpretations that the WWW of texts did not have; it was humans reading the texts who supplied the interpretations. The shift to a "data web" downplays that notion of comprehensible or cognitive content and it is that, that Hai Zhuge wants to put back. I salute the ambitious nature of his research goals that this book sets out.

Yorick Wilks

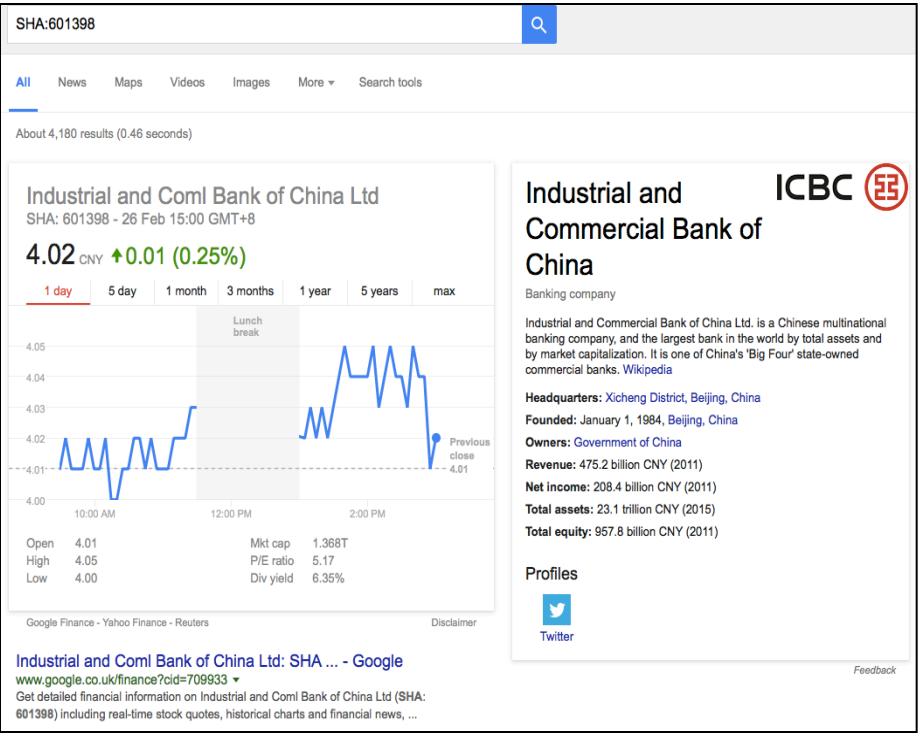
Application in Search Engine

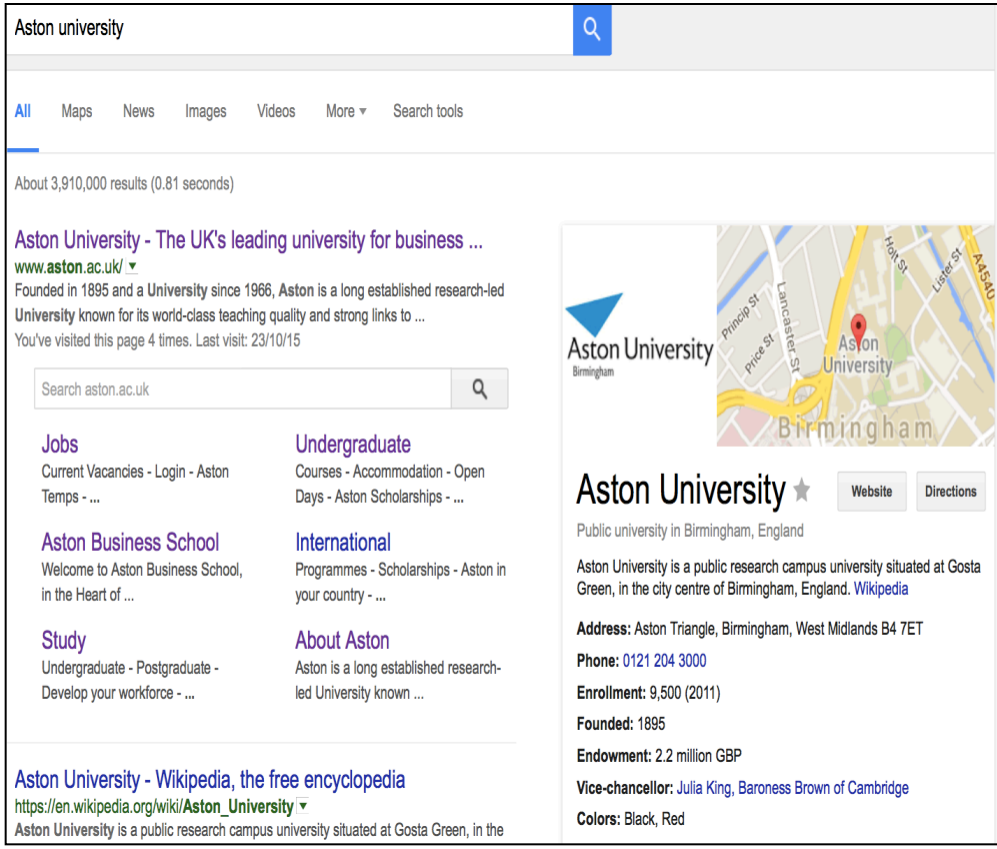


Extract Objects



Text + Chart + Map + List

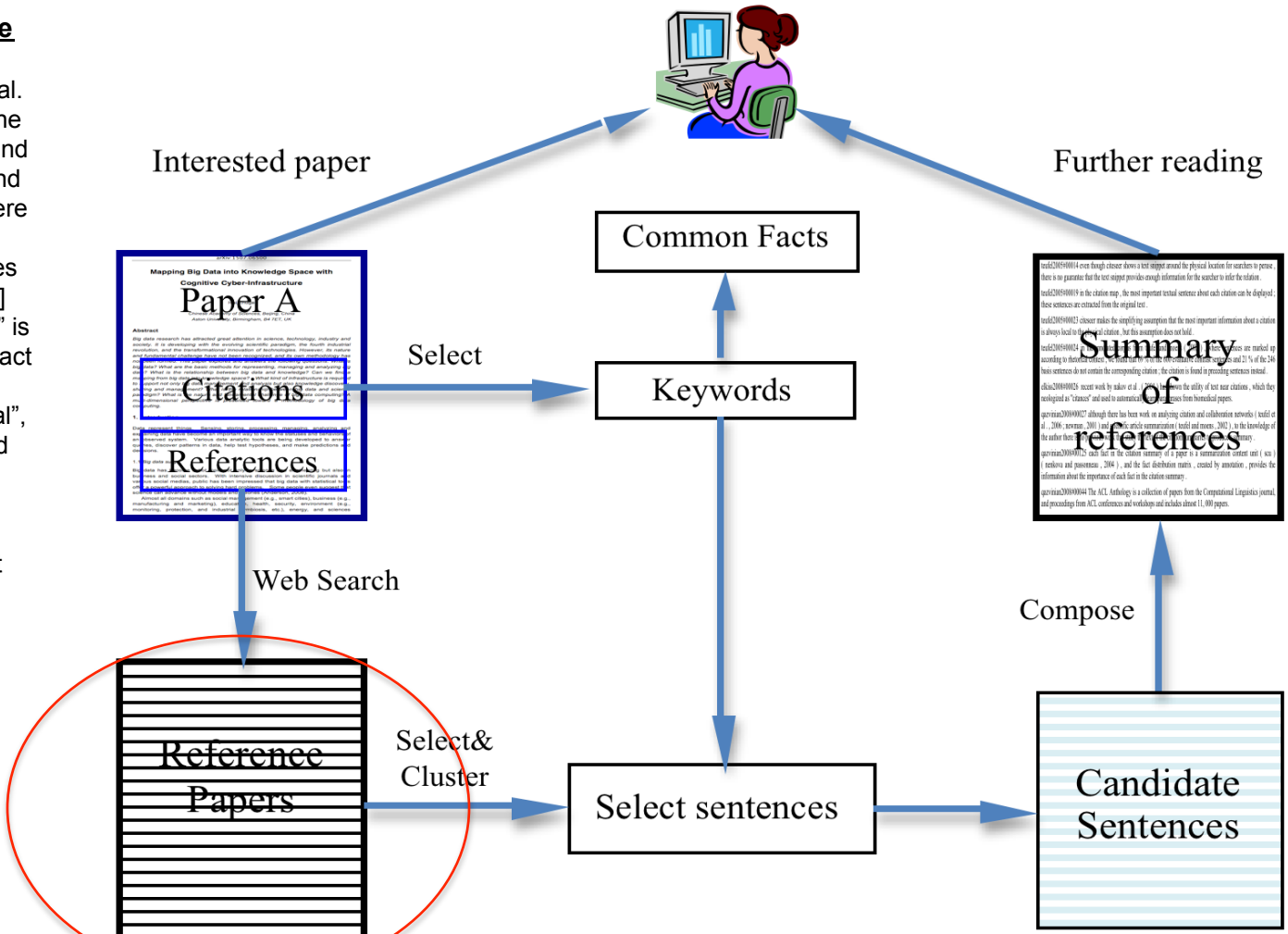




Consider Citation

Common Theme

"Information retrieval. Features such as the frequency of word and phrase, location, and rank of sentence were used to extract important sentences [Baxendale, 1958] [Edmundson, 1969]" is about the common fact indicated by "information retrieval", "frequency of word and phrase", "location", "rank sentence", and "extract important sentences".



Purpose: Summarize papers

Extend Application
Use Summary to Interpret

teufel2005#00014 even though citeseer shows a text snippet around the physical location for searchers to peruse , there is no guarantee that the text snippet provides enough information for the searcher to infer the relation .

teufel2005#00019 in the citation map , the most important textual sentence about each citation can be displayed ; these sentences are extracted from the original text .

teufel2005#00023 citeseer makes the simplifying assumption that the most important information about a citation is always local to the physical citation , but this assumption does not hold .

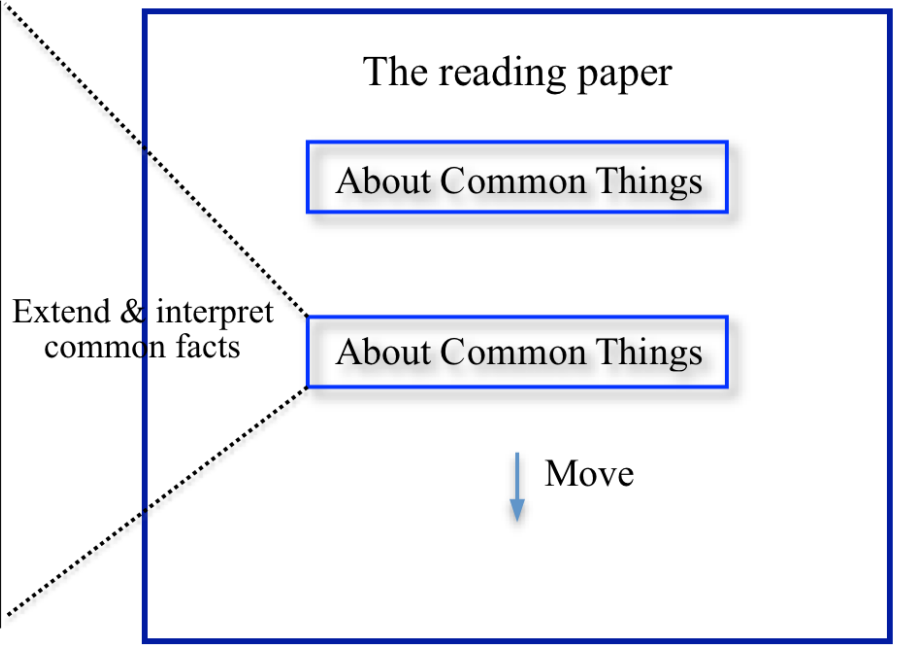
teufel2005#00024 in the annotated corpus from teufel and moens (2002) , where sentences are marked up according to rhetorical context , we found that 69 % of the 600 evaluative contrast sentences and 21 % of the 246 basis sentences do not contain the corresponding citation ; the citation is found in preceding sentences instead .

elkiss2008#00026 recent work by nakov et al . (2004) has shown the utility of text near citations , which they neologized as "citances" and used to automatically learn paraphrases from biomedical papers.

qazvinian2008#00027 although there has been work on analyzing citation and collaboration networks (teufel et al . , 2006 ; newman , 2001) and scientific article summarization (teufel and moens , 2002) , to the knowledge of the author there is no previous work that study the text of the citation summaries to produce a summary .

qazvinian2008#00125 each fact in the citation summary of a paper is a summarization content unit (scu) (nenkova and passonneau , 2004) , and the fact distribution matrix , created by annotation , provides the information about the importance of each fact in the citation summary .

qazvinian2008#00044 The ACL Anthology is a collection of papers from the Computational Linguistics journal, and proceedings from ACL conferences and workshops and includes almost 11, 000 papers.



Problems of Current Methods

- Too specific to apply to various forms of text
  - Words, phrases, and sentences — indicators rather than semantics
  - Interpretability is weak (no theory so far)
  - Knowledge needed for a good summarization is beyond text
  - Why can human use words beyond text?
- Neglect human behavior
  - Writers and readers experience with representation and summarization in lifetime.
  - Machines do not have the ability of experiencing
- Closed system
  - The process of summarization is closed
  - In cyberspace, reviews and comments are open, easily available, and valuable for improving summarization

Consider the Work System Self-Observation 1

- Minds emerge image-like representations when reading various texts
  - People talk and write according to the images
- *Can we realize a general summarization method through modeling human information processing mechanism*?

Self-Observation 2

- We can use own words when making summarization
- Human make summarization with the knowledge for understanding and the knowledge of using language
- *Can we explore the summarization problem from some basic structures and behaviors involved in language use and understanding*?

Self-Observation 3

- Human can make creative summarization with knowledge and reasoning
- Experience enables human to gain ability to observe and think through dimensions
  - People become more insightful when they can observe and think through more dimensions with deeper structure

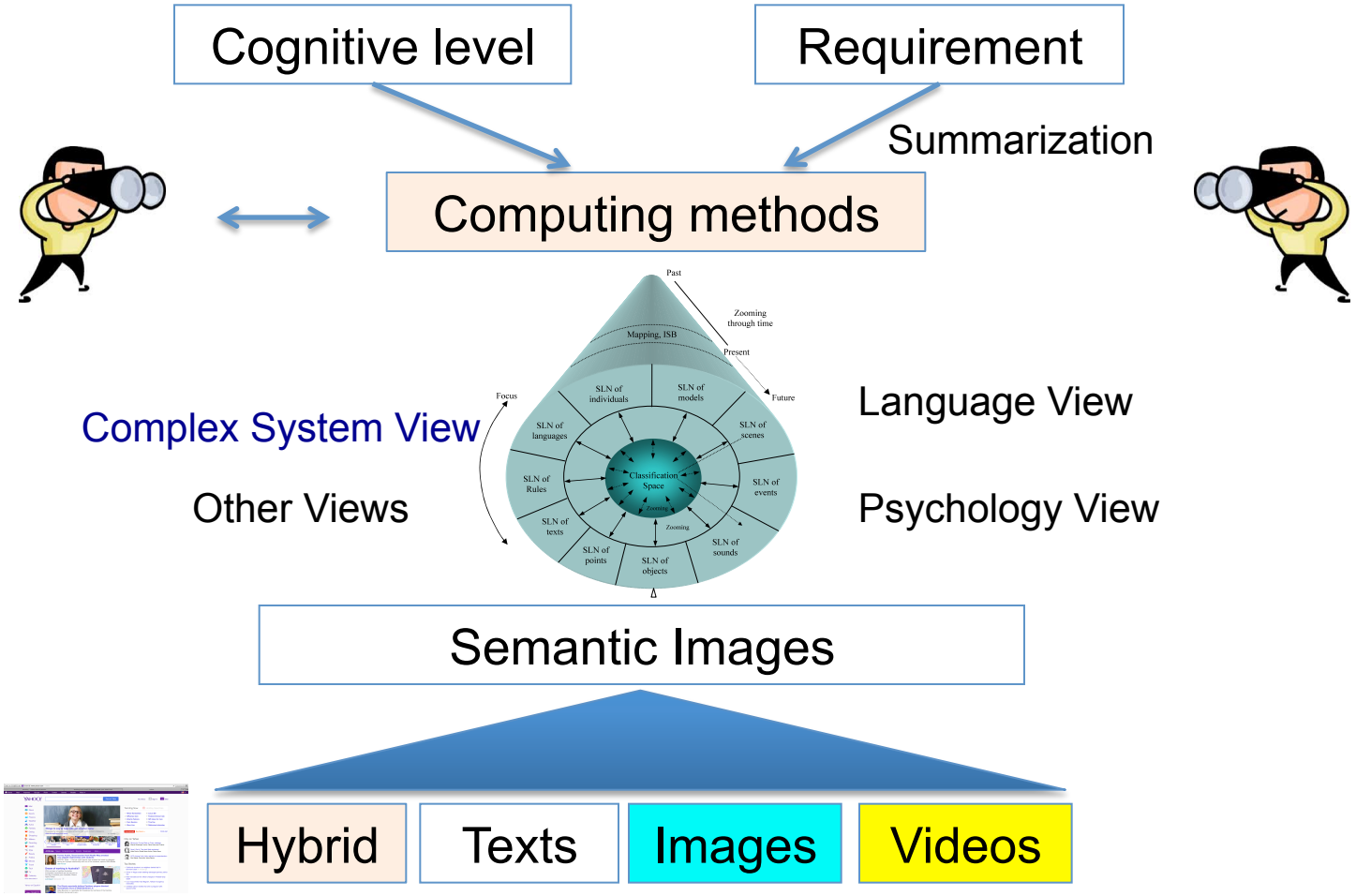
Self-Observation 4

- Human live in a Cyber-Physical-Social Space, evolving with sciences and technologies
- Can summarization be explored in a Cyber-Physical-Social Space to realize the Summarization of Things?

Identify Boundary

- Develop a computing process that can generate a piece of text from the original text according to predefined rules and constraints
- Develop an intelligent summarization system that aims at human-level summarization
- Philosophy of knowledge, intelligence, dimension, and cyber-physical society about summarization

General Solution: Think Lens

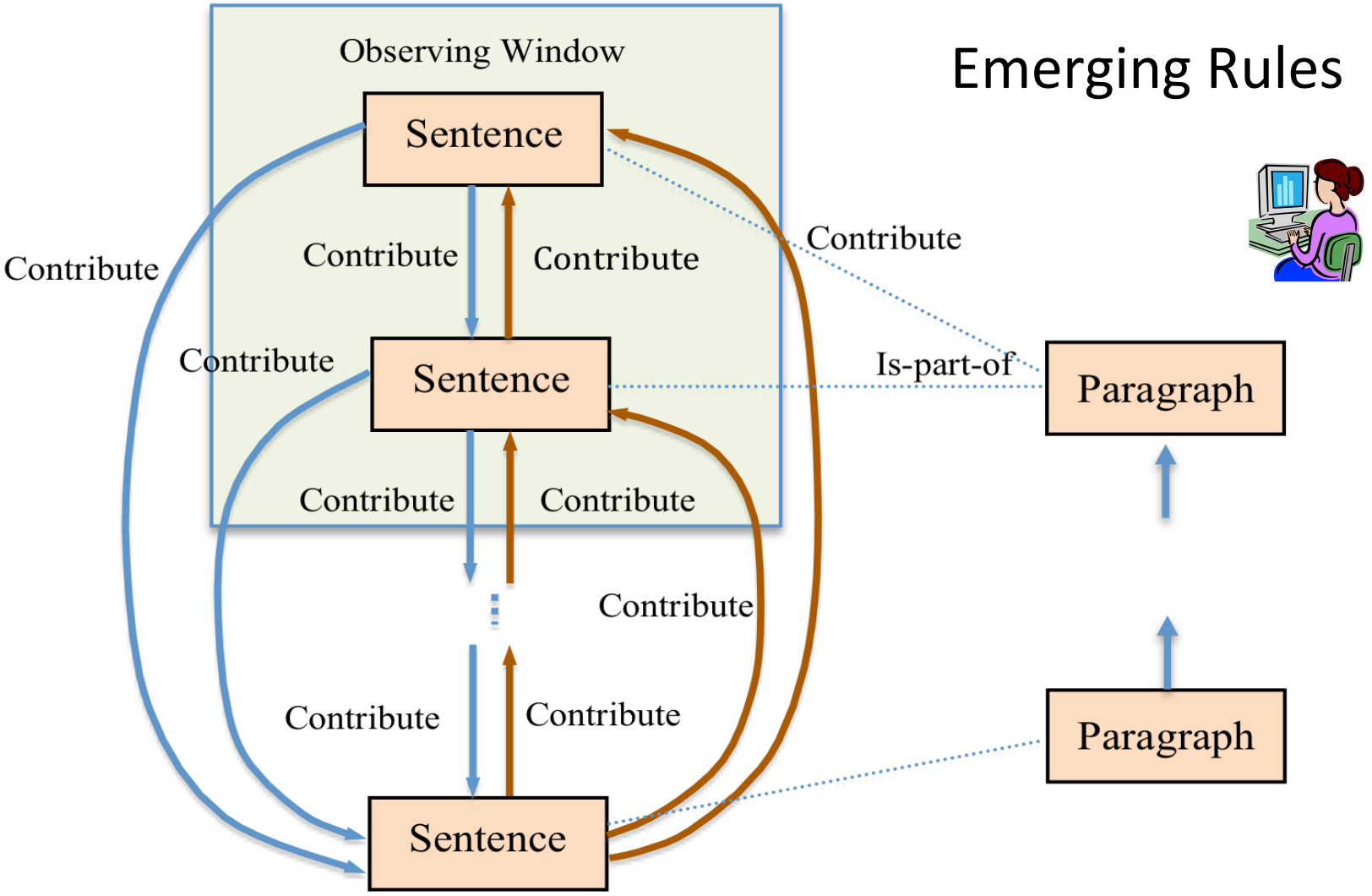


- Various Forms: Abstracts, CVs, Webpages, Posters, Slides

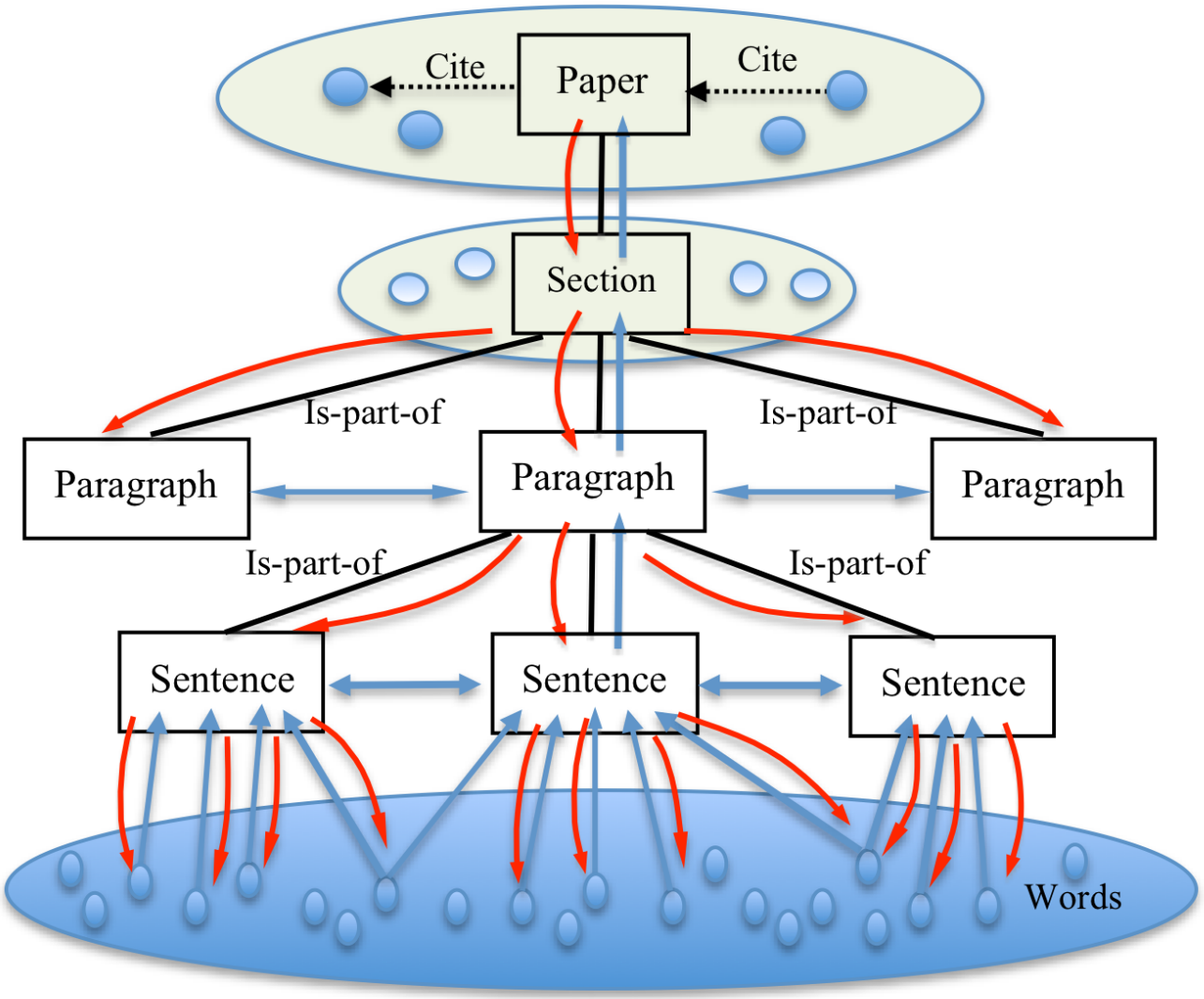
Zoom from Complex System's View

Emerging Structure

Emerging Structure



Reciprocal Relations



Ranking Sentences with the Reciprocal Relations

Sentences from abstract

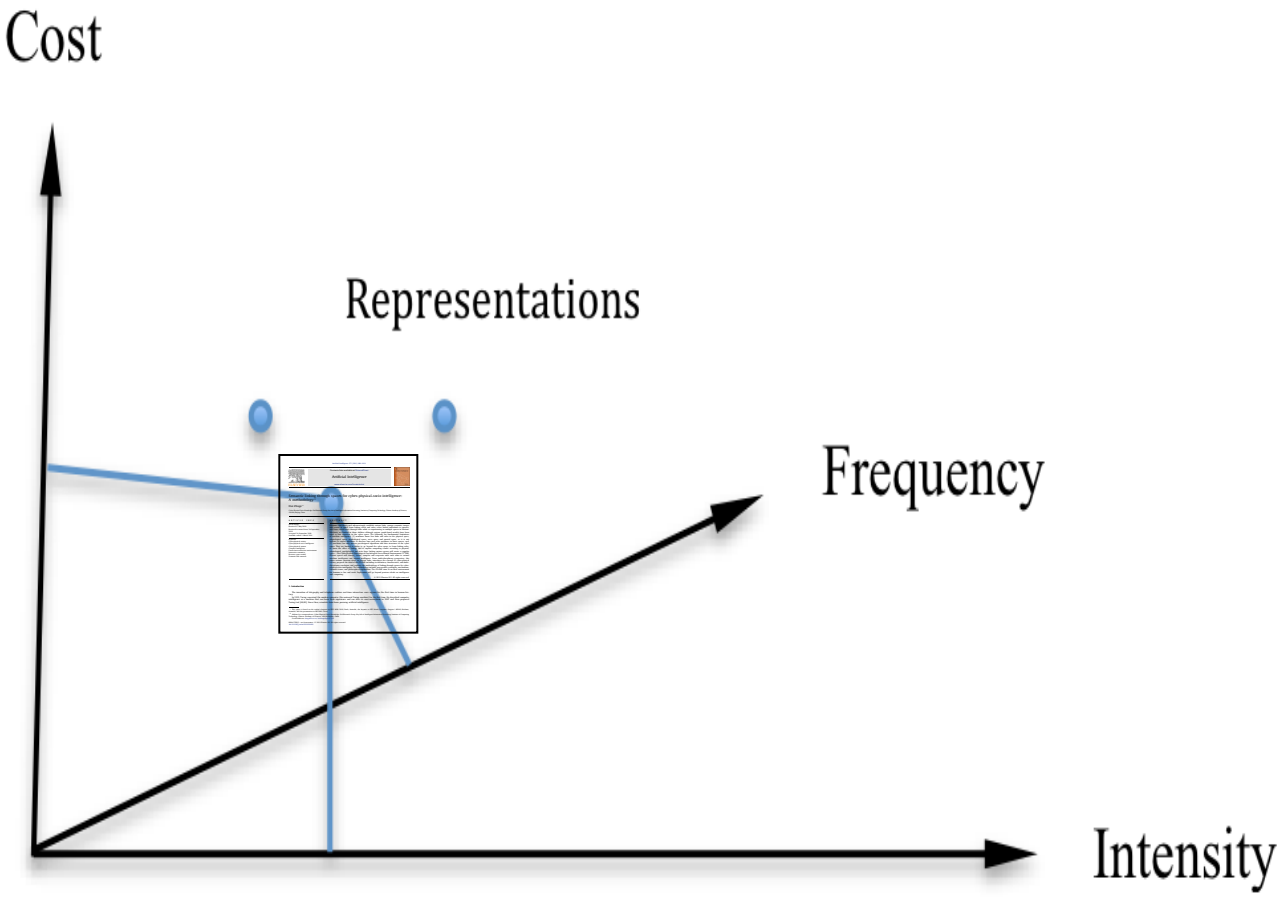
"*The basic viewpoints include:*
(1) a representation suitable for summarization should have a core, indicated by its intention and extension;
(2) summarization is an open process of various interactions, involved in various explicit and implicit citations;
(3) *and,the form of summary is diverse and summarization carries out through multiple dimensions.*"

Sentences selected from main text

The GS+SubPara Model:5/13

1. *e) Construct a semantic link network of pictures, tags and language representations in relevant texts as the summary.*
2. *A summarization system consisting of a multi-dimensional classification space of summary in form of semantic link network of pictures and language representations and a requirement space defined and managed by users. 1*
3. *The form of summary can be a semantic link network of texts, pictures, audios, and videos.*
4. *One semantic link network of texts can link to the other semantic link networks of texts to form a larger semantic link network through such relations as citation and coauthor. 0.5*
5. *The union of the semantic link network and the classification space forms a complex classification space used to represent and organize semantic images.*
6. *summary should use the core representations in the original representation. 1*
7. *The intention of representation is indicated by core representations and by citation from other representations. 0.5*
8. *The representation that has a direct link to the core representation is close-core representation. 1*
9. *A solution is to establish the semantic links between pictures, discover the communities of the semantic link network of pictures, select one picture to represent one community, and construct a network of the representative pictures.*
10. *A semantic link network of important concepts and relations can help readers quickly know the main cues and measures in representations.*
11. *The core representation renders the core idea of a representation. 1*
12. *Figure 8 depicts the idea of constructing a semantic link network of pictures and tags as a summary.*
13. *The semantic link network can be regarded as the map of the cyberspace and social space.*

Emerging Structure through Dimensions

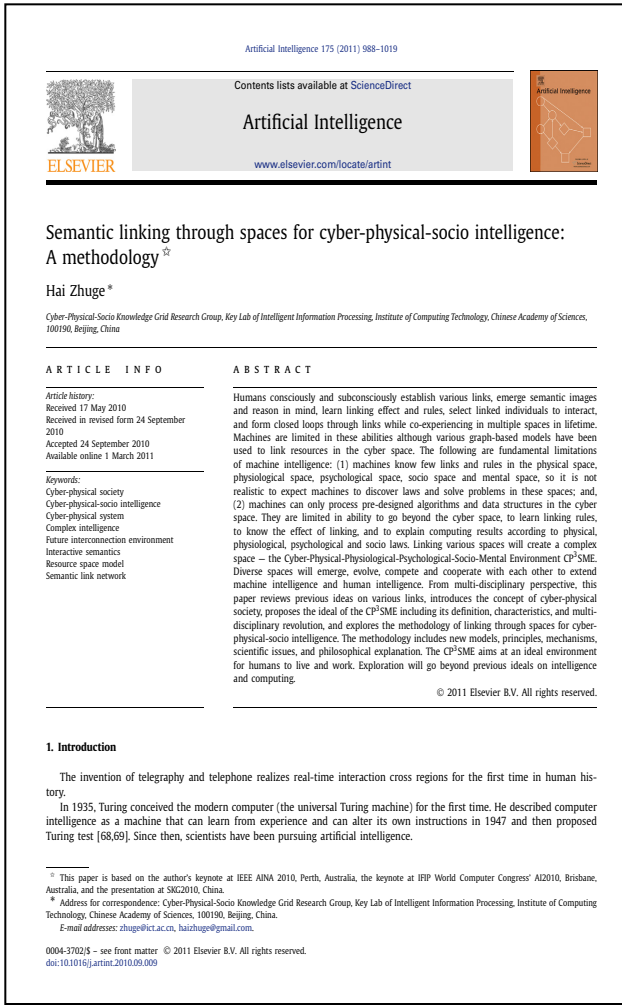


Cognitive level of Summarization

[1] The Rc (i) returns a set of the nodes that are the candidates of the root nodes in sentence i. 0
[2] We denote by rij the variable that is one if word i j is selected as a root of an extracting sentence subtree. 0
[3] We counted the number of sentences in the source document that each method used to generate a summary 5 5 Note that the number for the EDU method is not equal to selected textual units because a sentence in the source document may contain multiple EDUs. 0
[4] The sentence tree is a tree that has words as nodes and head modifier relationships between words obtained by the dependency parser as edges. 1
[5] We can build the nested tree by regarding each node of the document tree as a sentence tree. 1
[6] Our method jointly utilizes relations between sentences and relations between words, and extracts a rooted document subtree from a document tree whose nodes are arbitrary subtrees of the sentence tree. 1

Psychological View

- Goal assumption
  - A reader is to reach a goal through reading
- Coherence assumption
  - Integrity constraint
- Interpretation assumption
  - Readers try to interpret
    - Cause, effect, behaviors and states
  - Reading is mainly driven by generating and answering questions



Artificial Intelligence 175 (2011) 988–1019

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Semantic linking through spaces for cyber-physical-socio intelligence: A methodology ☆

Hai Zhuge *

Cyber-Physical-Socio Knowledge Grid Research Group, Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, 100190, Beijing, China

ARTICLE INFO

Article history:
Received 17 May 2010
Received in revised form 24 September 2010
Accepted 24 September 2010
Available online 1 March 2011

Keywords:
Cyber-physical society
Cyber-physical-socio intelligence
Cyber-physical system
Complex intelligence
Future interconnection environment
Interactive semantics
Resource space model
Semantic link network

ABSTRACT

Humans consciously and subconsciously establish various links, emerge semantic images and reason in mind, learn linking effect and rules, select linked individuals to interact, and form closed loops through links while co-experiencing in multiple spaces in lifetime. Machines are limited in these abilities although various graph-based models have been used to link resources in the cyber space. The following are fundamental limitations of machine intelligence: (1) machines know few links and rules in the physical space, physiological space, psychological space, socio space and mental space, so it is not realistic to expect machines to discover laws and solve problems in these spaces; and, (2) machines can only process pre-designed algorithms and data structures in the cyber space. They are limited in ability to go beyond the cyber space, to learn linking rules, to know the effect of linking, and to explain computing results according to physical, physiological, psychological and socio laws. Linking various spaces will create a complex space – the Cyber-Physical-Physiological-Psychological-Socio-Mental Environment CP^{3}SME. Diverse spaces will emerge, evolve, compete and cooperate with each other to extend machine intelligence and human intelligence. From multi-disciplinary perspective, this paper reviews previous ideas on various links, introduces the concept of cyber-physical society, proposes the ideal of the CP^{3}SME including its definition, characteristics, and multi-disciplinary revolution, and explores the methodology of linking through spaces for cyber-physical-socio intelligence. The methodology includes new models, principles, mechanisms, scientific issues, and philosophical explanation. The CP^{3}SME aims at an ideal environment for humans to live and work. Exploration will go beyond previous ideals on intelligence and computing.

1. Introduction

The invention of telegraphy and telephone realizes real-time interaction cross regions for the first time in human history.

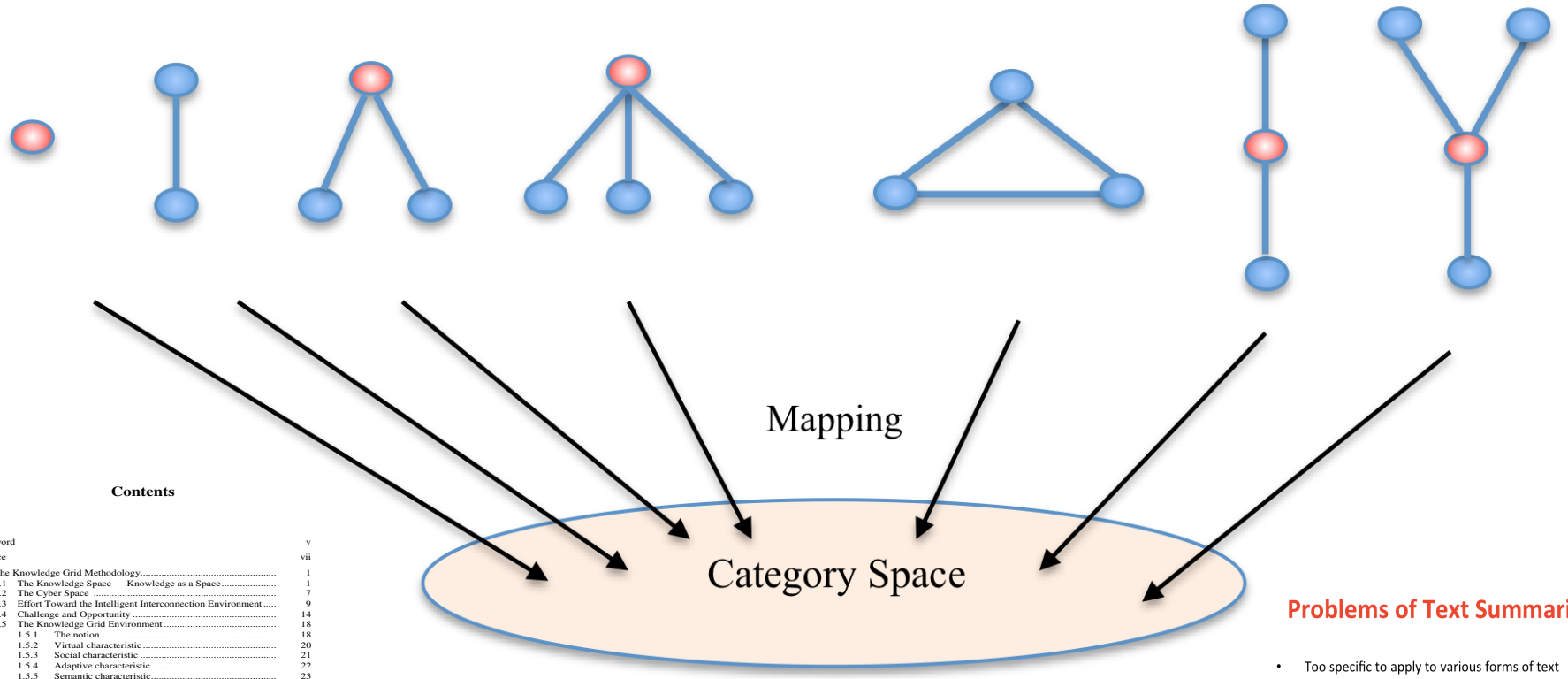
In 1935, Turing conceived the modern computer (the universal Turing machine) for the first time. He described computer intelligence as a machine that can learn from experience and can alter its own instructions in 1947 and then proposed Turing test [68,69]. Since then, scientists have been pursuing artificial intelligence.

☆ This paper is based on the author's keynote at IEEE AINA 2010, Perth, Australia, the keynote at IFIP World Computer Congress' AI2010, Brisbane, Australia, and the presentation at SKG2010, China.

* Address for correspondence: Cyber-Physical-Socio Knowledge Grid Research Group, Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, 100190, Beijing, China.
E-mail addresses: zhuge@ict.ac.cn, haizhuge@gmail.com.

0004-3702/$ – see front matter
doi:10.1016/j.artint.2010.09.009

Patterns Make Representation Decomposable

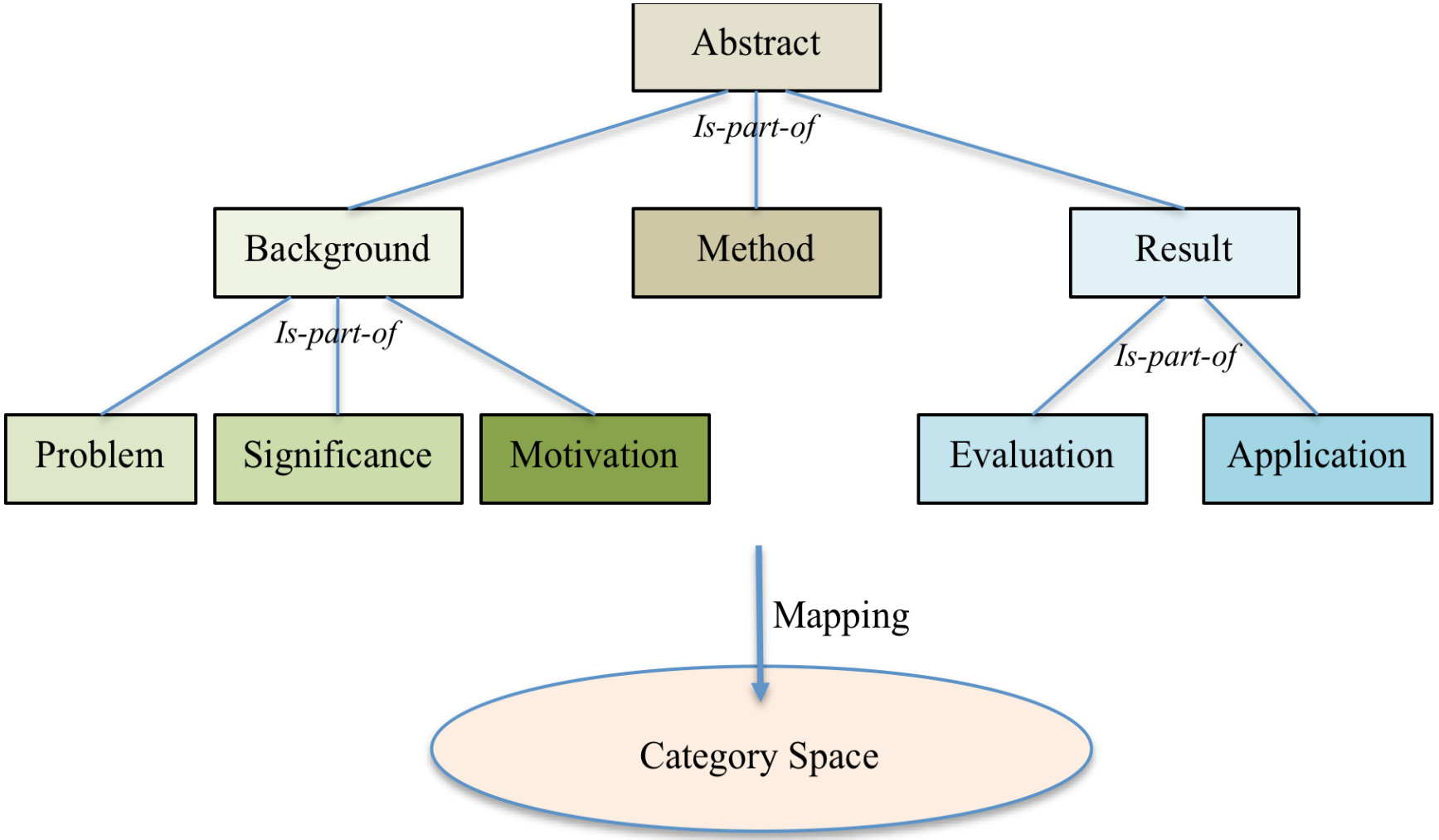


Contents

Problems of Text Summarization

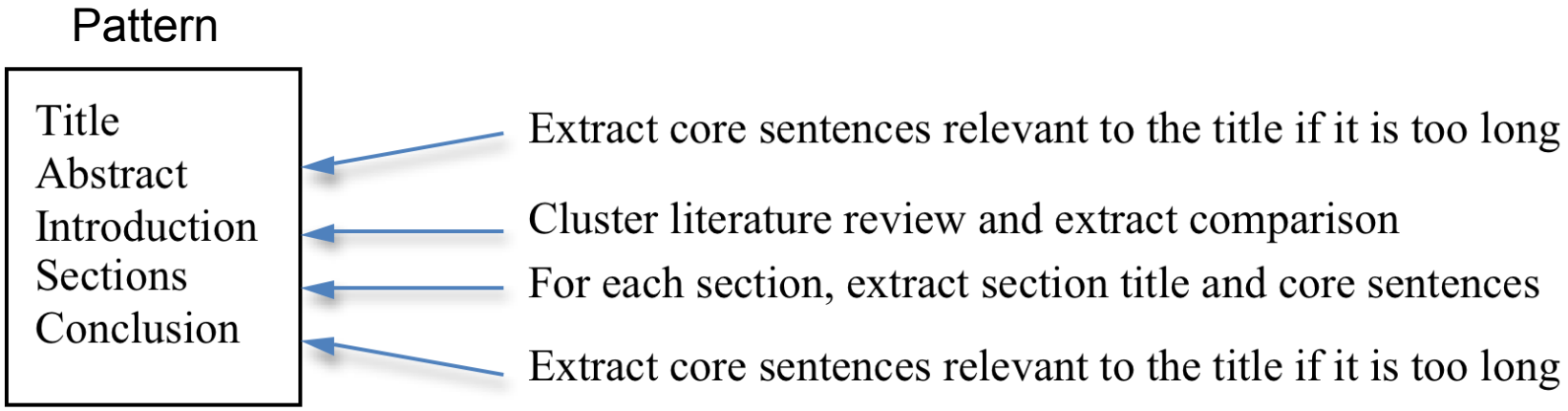
- Too specific to apply to various forms of text
  - Words, phrases, and sentences — indicators rather than semantics
  - Interpretability is weak (no theory)
  - Knowledge needed for a good summarization is beyond text
  - Why can human use words beyond text?
- Neglect human behavior
  - Writers and readers experience with representation and summarization in lifetime.
  - Machines do not have the ability of experiencing
- Closed system
  - The process of summarization is closed
  - In cyberspace, reviews and comments are open, easily available, and valuable for improving summarization

The Pattern of Abstract



Pattern and Understanding

- Recognizing various patterns is a part of human intelligence
- Patterns help representation & understanding



Problems of Text Summarization

- Too specific to apply to various forms of text
  - Words, phrases, and sentences — indicators rather than semantics
  - Interpretability is weak (no theory)
  - Knowledge needed for a good summarization is beyond text
  - Why can human use words beyond text?
- Neglect human behavior
  - Writers and readers experience with representation and summarization in lifetime.
  - Machines do not have the ability of experiencing
- Closed system
  - The process of summarization is closed
  - In cyberspace, reviews and comments are open, easily available, and valuable for improving summarization

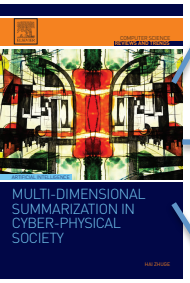
Proposition

The long-term behaviors more rely on complex patterns while the short-term behaviors more rely on the basic patterns.

Generalization: Dimensions

- Fundamental problem of information service:
  - the modeling and maintenance of a multi-dimensional interest space for various users
  - a multi-dimensional category space for various resources
  - the mapping from the interest space into the resource space.

Dimensions of Summarization



Publisher

Expert

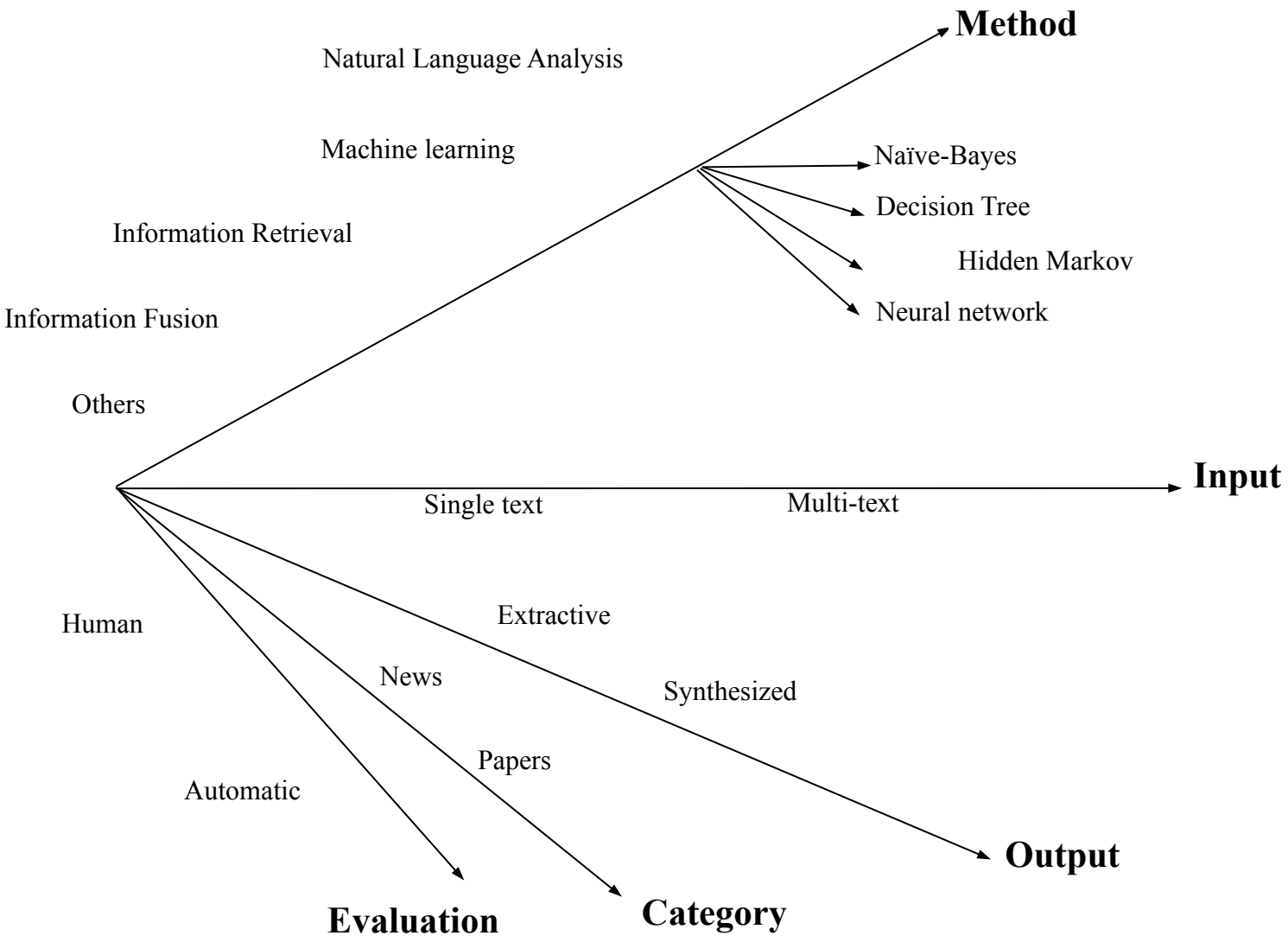
Author

This insightful resource presents a methodology for general summarization in cyber-physical-social space through a multi-dimensional lens of semantic computing. It transforms the paradigm of summarization research and deepens people's understanding on semantics, dimension, knowledge, and computing.

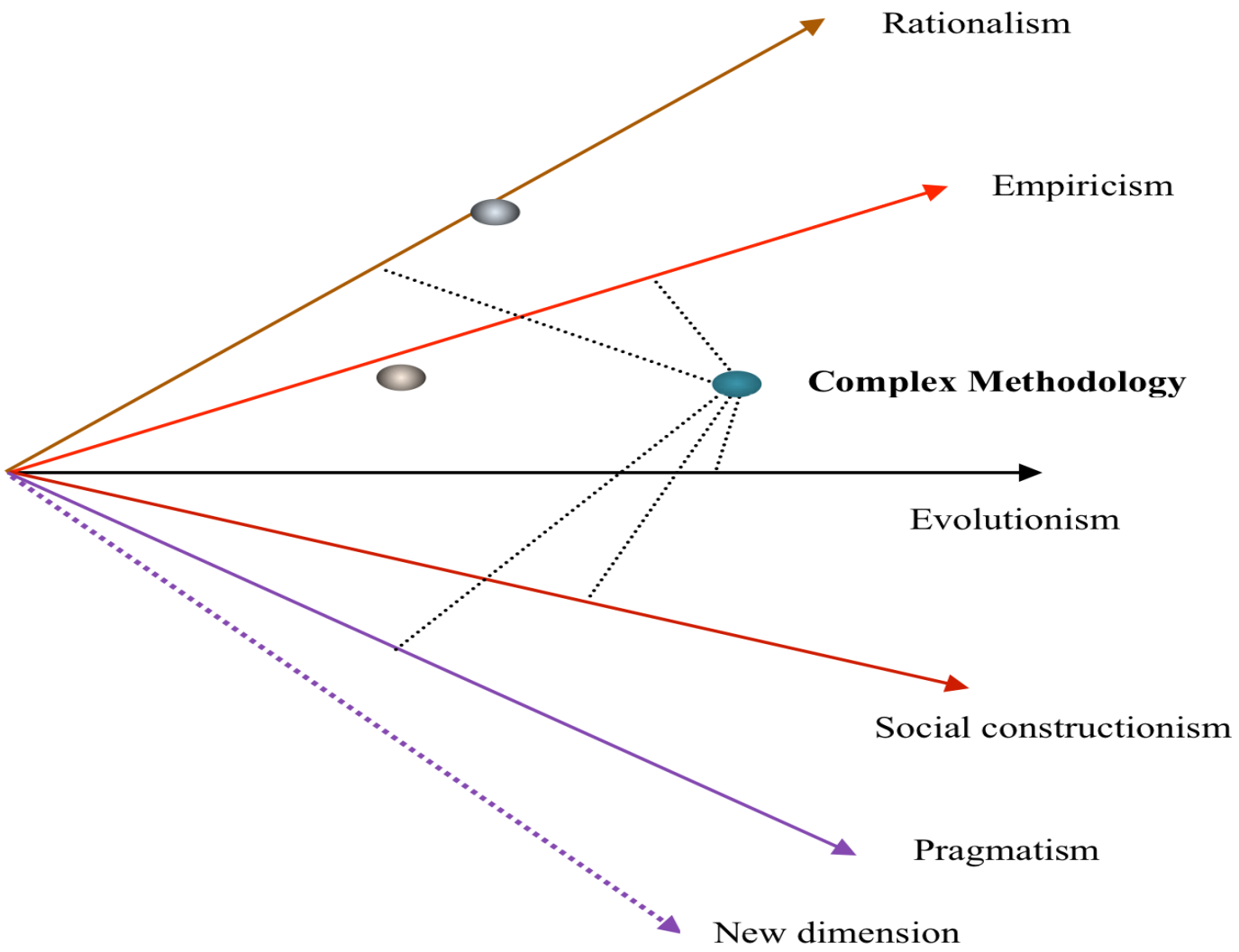
Within the wider Semantic Web research project, I think Hai Zhuge is trying to do something bold and important: first, by asking how can big data — and all that that movement now entails — be made to yield up representations that that give us some understanding of the content implicit in the data, but is not essentially textual. This is his Resource Space Model, a multi-dimensional category space that uniformly represents human understanding and the patterns in real and abstract objects. Secondly, he is asking how can we have a common representation for the content of different modalities: not just speech and text but images and 3D objects, the long-awaited Internet of Things. He then takes summarization as a reasonably well-understood task environment where the notion of summarization beyond the textual — the standard case so far — can be explored and ultimately tested. I think he is asking serious and fundamental questions in this work and proposing bold research solutions. This is an important issue in the history of idea development of the semantic web where many have transitioned seamlessly from the term "semantic web" to the "data web", without always thinking what that entails. For example, the whole motivation of the semantic web was to have a web of understandable meanings or interpretations that the WWW of texts did not have; it was humans reading the texts who supplied the interpretations. The shift to a "data web" downplays that notion of comprehensible or cognitive content and it is that that Hai Zhuge wants to put back. I salute the ambitious nature of his research goals that this book sets out. -- Foreword by **Yorick Wilks**.

Text summarization has been studied for over a half century, but traditional methods process texts empirically and neglect the fundamental characteristics and principles of language use and understanding. Automatic summarization is a desirable technique for processing big data. This reference summarizes previous text summarization approaches in a multi-dimensional category space, introduces a multi-dimensional methodology for research and development, unveils the basic characteristics and principles of language use and understanding, investigates some fundamental mechanisms of summarization, studies dimensions on representations, and proposes a multi-dimensional evaluation mechanism. Investigation extends to incorporating pictures into summary and to the summarization of videos, graphs, and pictures, and converges to a general summarization method. Furthermore, some basic behaviors of summarization are studied in the complex cyber-physical-social space. Finally, a creative summarization mechanism is proposed as an effort toward the creative summarization of things, which is an open process of interactions among physical objects, data, people, and systems in cyber-physical-social space through a multi-dimensional lens of semantic computing. The author's insights can inspire research and development of many computing areas.

Dimensions on Methods



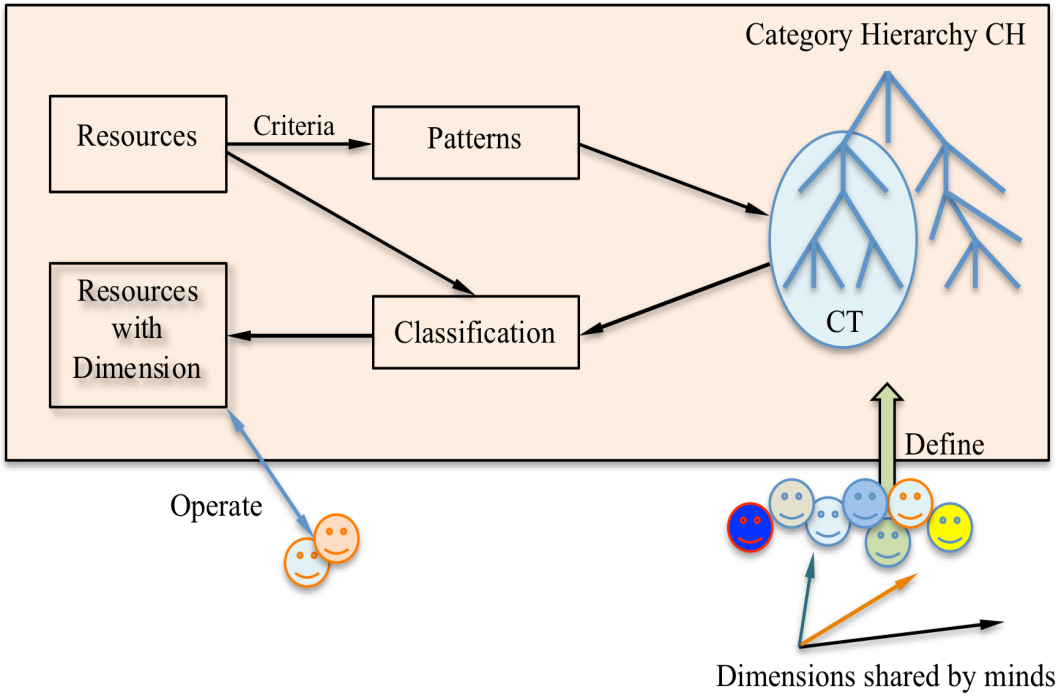
Multi-Dimensional Methodology



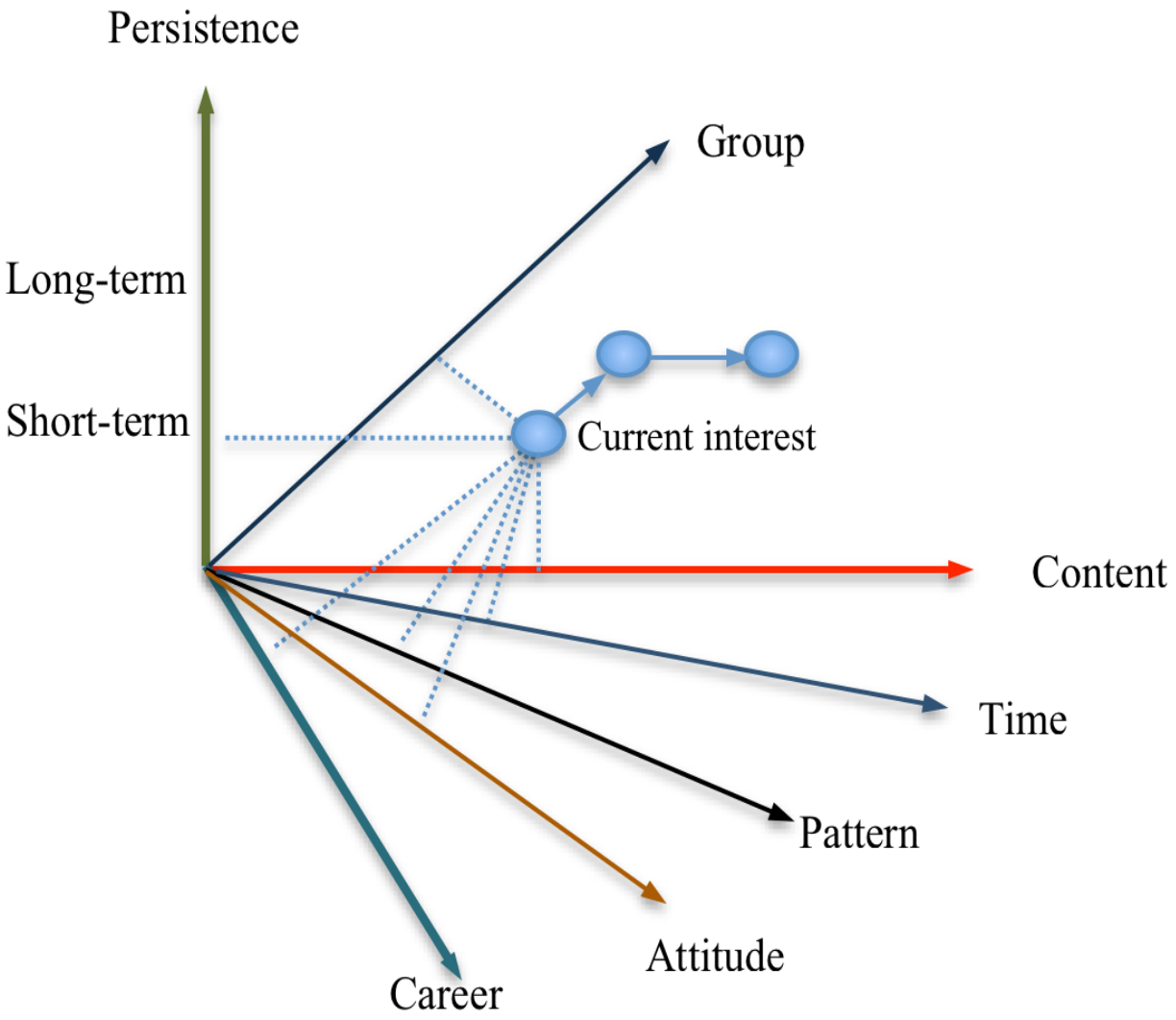
How to discover dimensions?

Problem Definition. Given a set of resources R and a category hierarchy CH, find a set of criteria to extract a set of feature representations W from R, and then find a category tree CT within CH according to W such that R can be classified by CT.

- If pattern of CH is stable
  - Use it to evaluate the patterns in resources
- If pattern in resources is stable
  - Use it to evaluate the pattern of CH

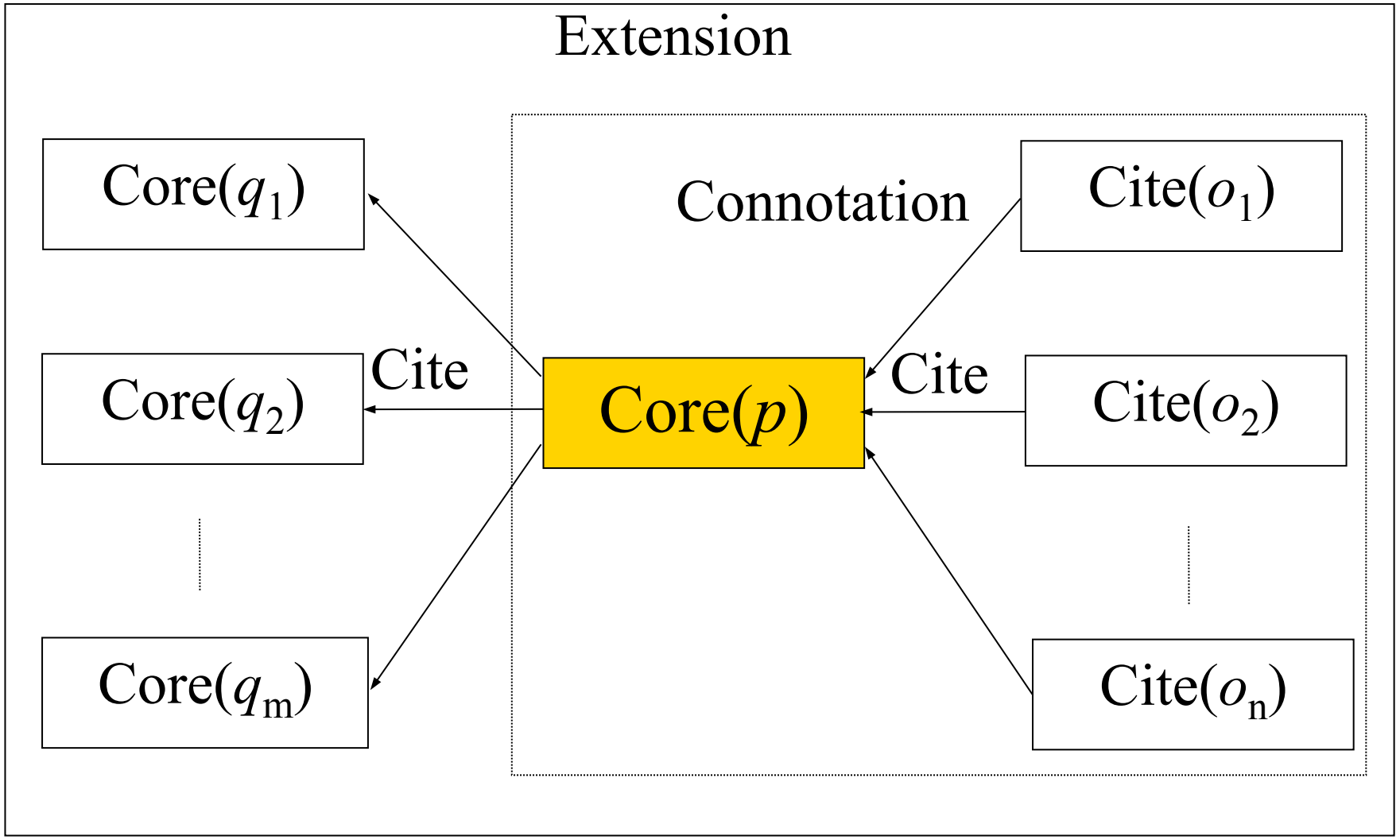


Multi-Dimensional Interest Space

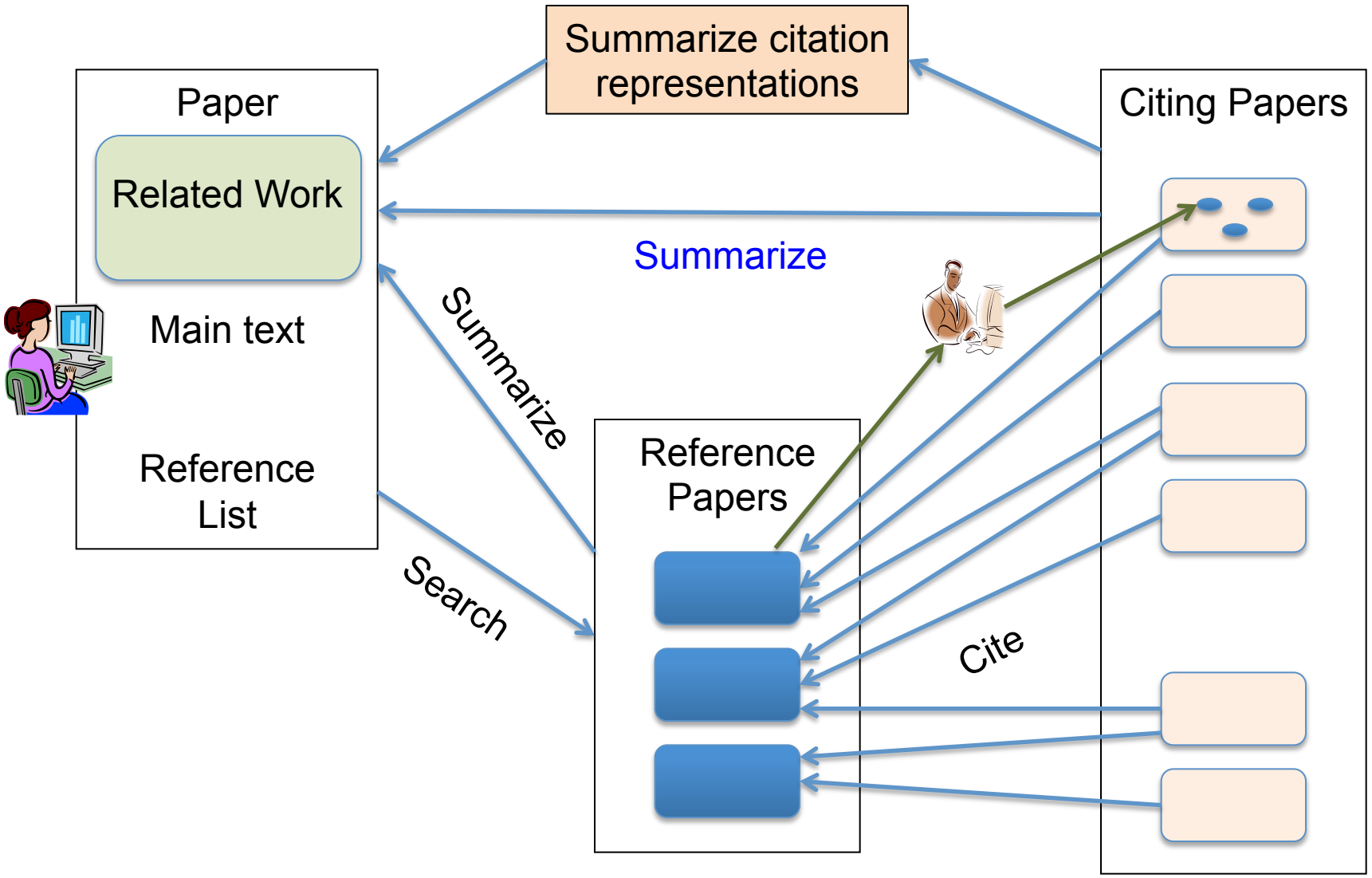


Rediscovering

Representing with Citation



Example: Generating Related Work



Summarization and Citation

- Summarization is generated in advanced development stage of language when complex structure emerges
  - Citation is a basic approach to constructing a complex text
- **Generally, summarization is a process of forming explicit and implicit citations**

General Citation

- An individual selection of relevant representations for explaining, evidencing, and complementing a representation
  - either explicitly or implicitly
  - selections of individuals form social selection
  - ranks texts and evolves the network of citation

Extend the Scope

Linking People and Society

Semantic Link Network of people

Read

Write

Semantic Link Network of texts

Is-part-of

Semantic Link Network within text

Interpret

Summary

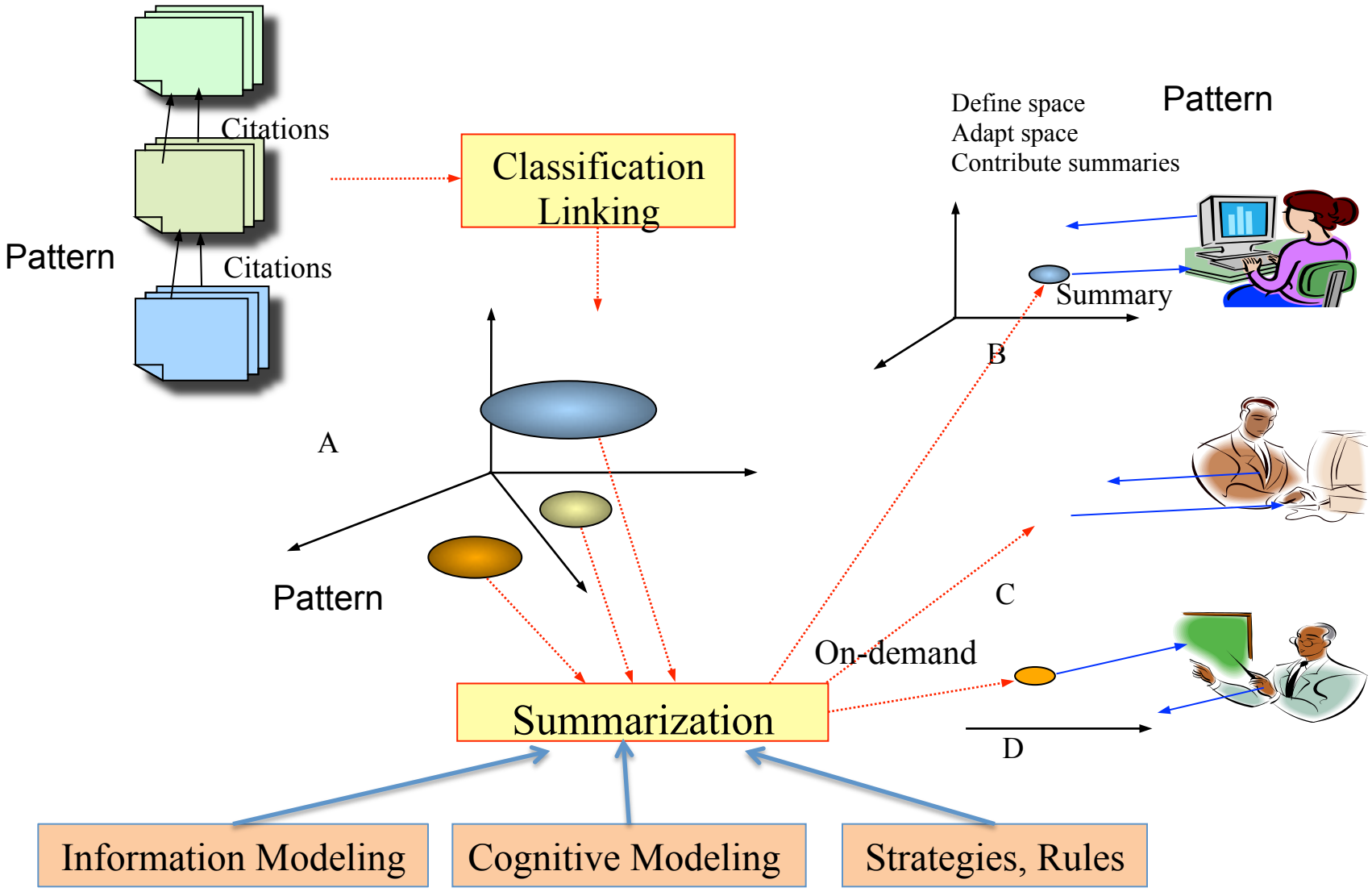
Text

Summarization

Physical Dimension

- **Efficiency**
  - Humans behave consciously or unconsciously with lower consumption of energy and time
- **Locality**
  - Human psychological system tends to behave in a smaller space and with less time

Summarization with Dimensions



Zooming Out
Paradigm Shift

- Application driven
  - Application (e.g., scale requirement of summary) ⇒
  - Method (find a method) ⇒ Evaluation (with data)

Data driven

Data (texts)

Representing (e.g., reading, writing, and citing)

Information modeling

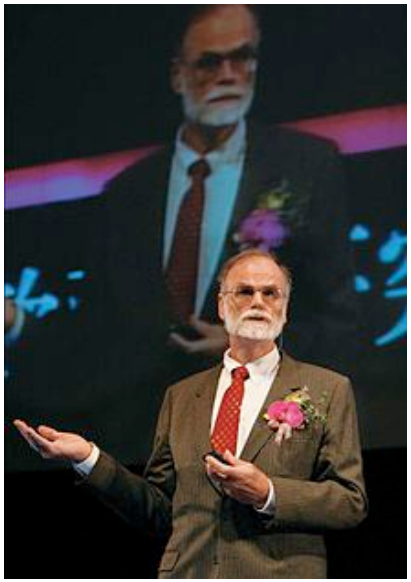
Cognitive modeling

Method: Apply knowledge through information modeling to solve any problem

Knowledge modeling

One of Jim Gray's Dreams (1999)

- Summarization as expert
  - As a function of World Memex
  - Precise
  - Quick
  - Record everything
  - QA



Studying the transformation of the research paradigm in the summarization area with cross-disciplinary inspiration provides a case study for exploring the methodology of data science and cyber-physical society

References

- H. Zhuge, Multi-Dimensional Summarization in Cyber-Physical Society, Morgan Kaufmann, 2016.
- H.Zhuge, The Knowledge Grid — Toward Cyber-Physical Society, World Scientific Publishing Co, 2012 (2nd Edition).
- H.Zhuge and Y.Xing, Probabilistic Resource Space Model for Managing Resources in Cyber-Physical Society, *IEEE Transactions on Service Computing*, 5(3)(2012)404-421.
- H.Zhuge, Semantic linking through spaces for cyber-physical-socio intelligence: A methodology. *Artificial Intelligence,* 175(2011)988–1019.
- H.Zhuge, Interactive Semantics. *Artificial Intelligence, 174* (2010)190–204.
- H.Zhuge, The Web Resource Space Model, Springer, 2008.
- H.Zhuge, The Future Interconnection Environment, *IEEE Computer*, 38(4) (2005)27-33.
- H.Zhuge, Communities and Emerging Semantics in Semantic Link Network: Discover and learning, *IEEE Transactions on Data and Knowledge Engineering*, 21(6)(2009)785-799.