Constructing Virtual Documents for Ontology Matching

Yuzhong Qu, Wei Hu, Gong Cheng

Southeast University, China

Outline

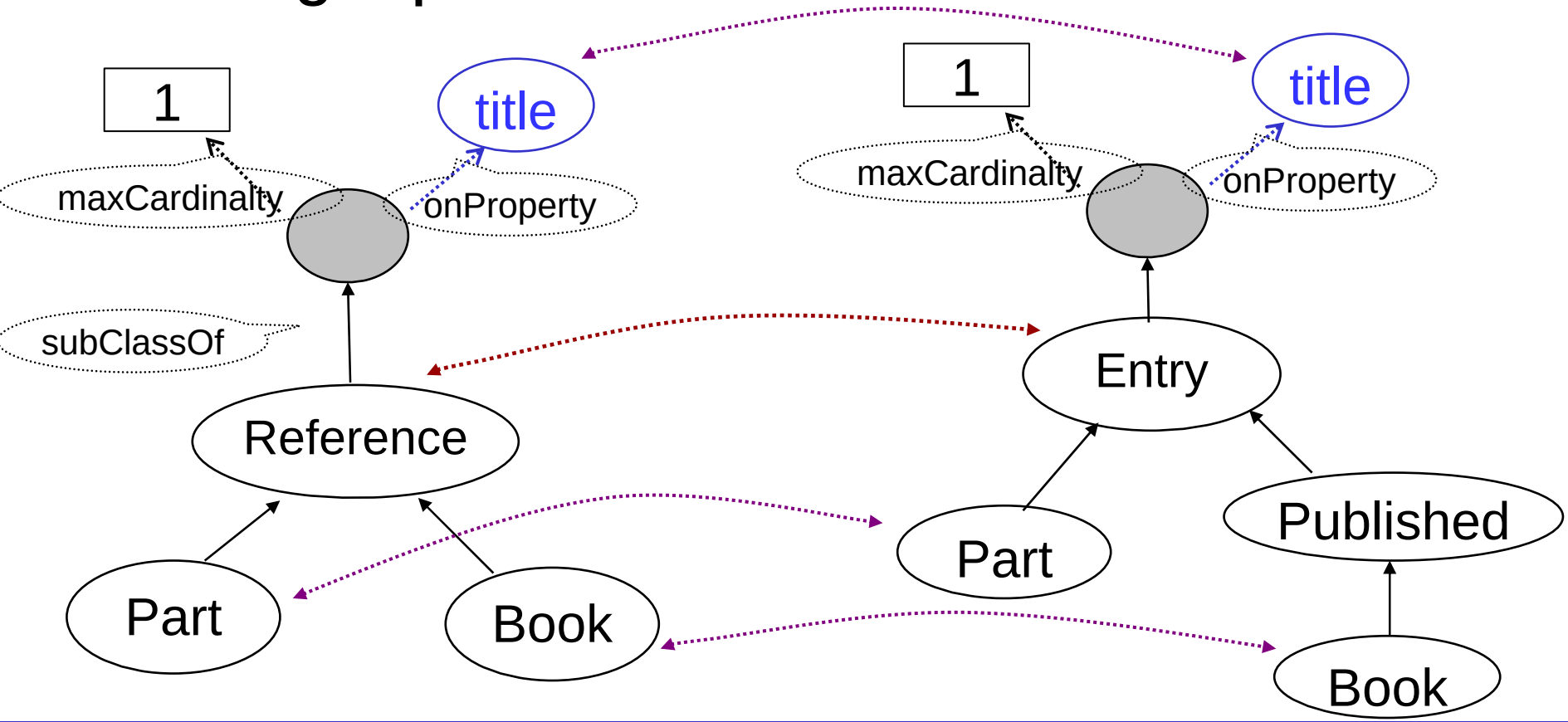
- Introduction
- Investigation on Linguistic Matching
- Main Idea of V-Doc Approach
- Formulation of Virtual Documents
- Experiments
- Concluding Remarks

Introduction

- Ontology
  - A key to SW (Semantic Web)
  - More ontologies are written in RDFS, OWL
  - It's not unusual:
    - ✓Multiple ontologies for overlapped domains (Diversity of Voc)
- Ontology Matching
  - Important to SW applications, but difficult
  - Inherent difficulty
    - ✓The complex nature of RDF graph
    - ✓The heterogeneity in structures and linguistics (labels)

Introduction (Example)

- bibliographic references VS bibTeX



Introduction (Cont.)

- Techniques
  - Linguistic matching: string comparison, synonym
  - Structural matching: “similarity propagation”
    - ✓ Originated from Cupid and Similarity Flooding (match DB schema)
- Algorithms and tools
  - Cupid, OLA, ASCO, HCONE-merge, SCM, GLUE, S-Match
  - PROMPT, QOM, Falcon-AO
- “Standard" tests
  - OAEI 2005 (KCAP2005), EON 2004, and I3CON 2003

Introduction (Cont.)

- Though the formulation of structural matching is a key feature of a matching approach
- Ontology matching should ground on linguistic matching
- Main focus: Linguistic matching for ontologies

Investigation on linguistic matching(1)

- Label/name comparison is exploited well
  - Levenshtein's edit distance, I-Sub
- Descriptions (comments, annotations)
  - Are used in some tools
  - NOT yet been exploited very well
- Neighboring information
  - Is partially used in some tools
  - Need to be explored systematically

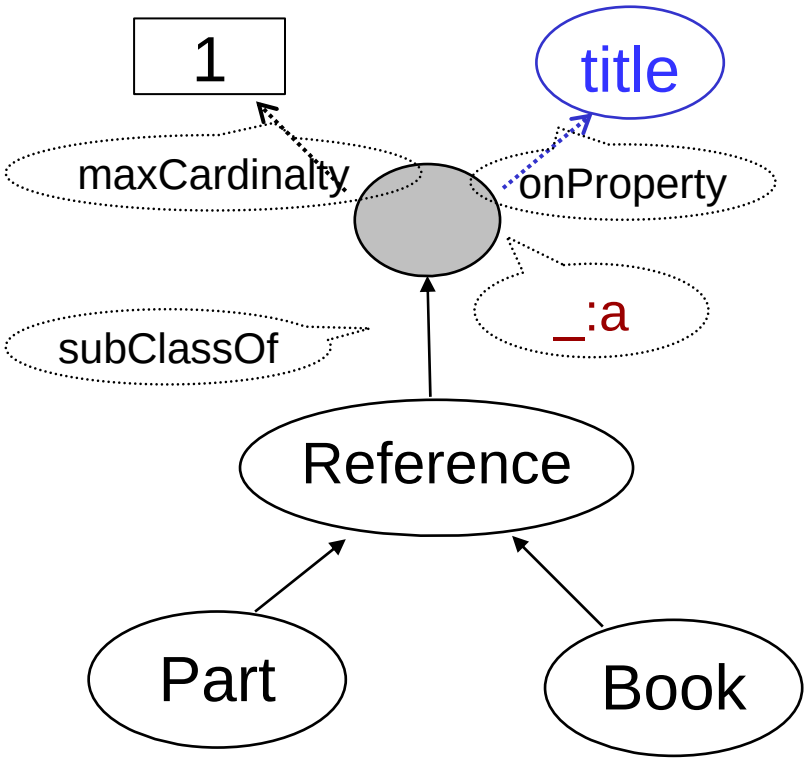
Investigation on linguistic matching(2)

- Looking up synonym (WordNet) is time-consuming
  - OLA in OAEI 2005 contest
    - ✓ The string distance methods have better performances and are also much more efficient than the ones using WordNet-based computation.
  - Also reported by the experience of ASCO
    - ✓ Integration of WordNet in the calculation of description similarity may not be valuable and cost much time.
  - Our own experimental results (shown later)
    - ✓ WordNet-based computation faces the problem of efficiency and accuracy in some cases.

Main Idea of V-Doc Approach (1)

- Encode the intended meaning of named nodes in OWL/RDF ontologies via virtual documents
- Take the similarity between VDs (Cosine, TF/IDF) as the similarity between named nodes
- The virtual document for each named node (URIref)
  - Is a collection of weighted words
  - Includes not only local descriptions but also neighboring information.

Main Idea of V-Doc Approach (2)



- VD(ex1:Reference)
  - Local Description
  - Des(ex1:Part)
  - Des(ex1:Book)
  - Des(_:a)

Formulation of Virtual Documents(1)

- The (local) description of a named node

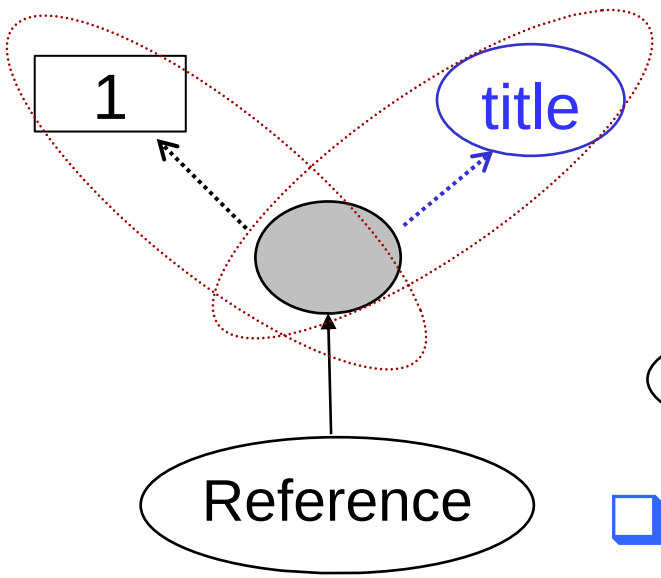
$$
\begin{aligned}
Des(e) \quad = \quad & \alpha_1 * \text{collection of words in the local name of } e \\
& + \alpha_2 * \text{collection of words in the rdfs:label of } e \\
& + \alpha_3 * \text{collection of words in the rdfs:comment of } e \\
& + \alpha_4 * \text{collection of words in other annotations of } e
\end{aligned}
$$

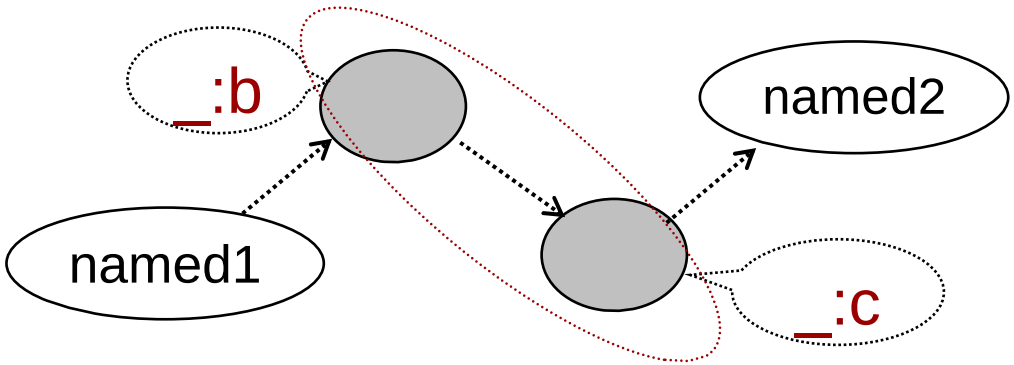
Formulation of Virtual Documents(2)

- The description of a blank node

$$Des_1(b) = \beta * \sum_{sub(s)=b} Des(pre(s)) + Des(obj(s))$$

$$Des_{k+1}(b) = Des_k(b) + \beta * \sum_{\substack{sub(s)=b \\ obj(s)\in B}} Des_k(obj(s)) \qquad (k \geq 1)$$





- $Des_2(\_:b) = \beta \; Des_1(\_:c) + \ldots$

Formulation of Virtual Documents(3)

- The virtual document of a named node

$$VD(e) = Des(e) + \gamma_1 * \sum_{e' \in SN(e)} Des(e') + \gamma_2 * \sum_{e' \in PN(e)} Des(e') + \gamma_3 * \sum_{e' \in ON(e)} Des(e')$$

- *SN*(*e*): subject neighboring
  - The nodes that occur in triples with e as the subject
- *PN*(*e*): predicate neighboring
- *ON*(*e*): object neighboring

Formulation of Virtual Documents(4)

- Examples of Virtual documents
  - VD(ex1:Reference)=
    - ✓ {(reference, 1.46), (title, 0.027), (part, 0.005), (book, 0.004), …}
  - VD(ex2:Entry)=
    - ✓ {(entry, 1.66), (title, 0.031), (part, 0.005), (book, 0.008), (publish,0.007), …}
- Similarity(ex1:Reference, ex2:Entry)=0.284
  - Cosine, tfidf

Experiments — Setting(1)

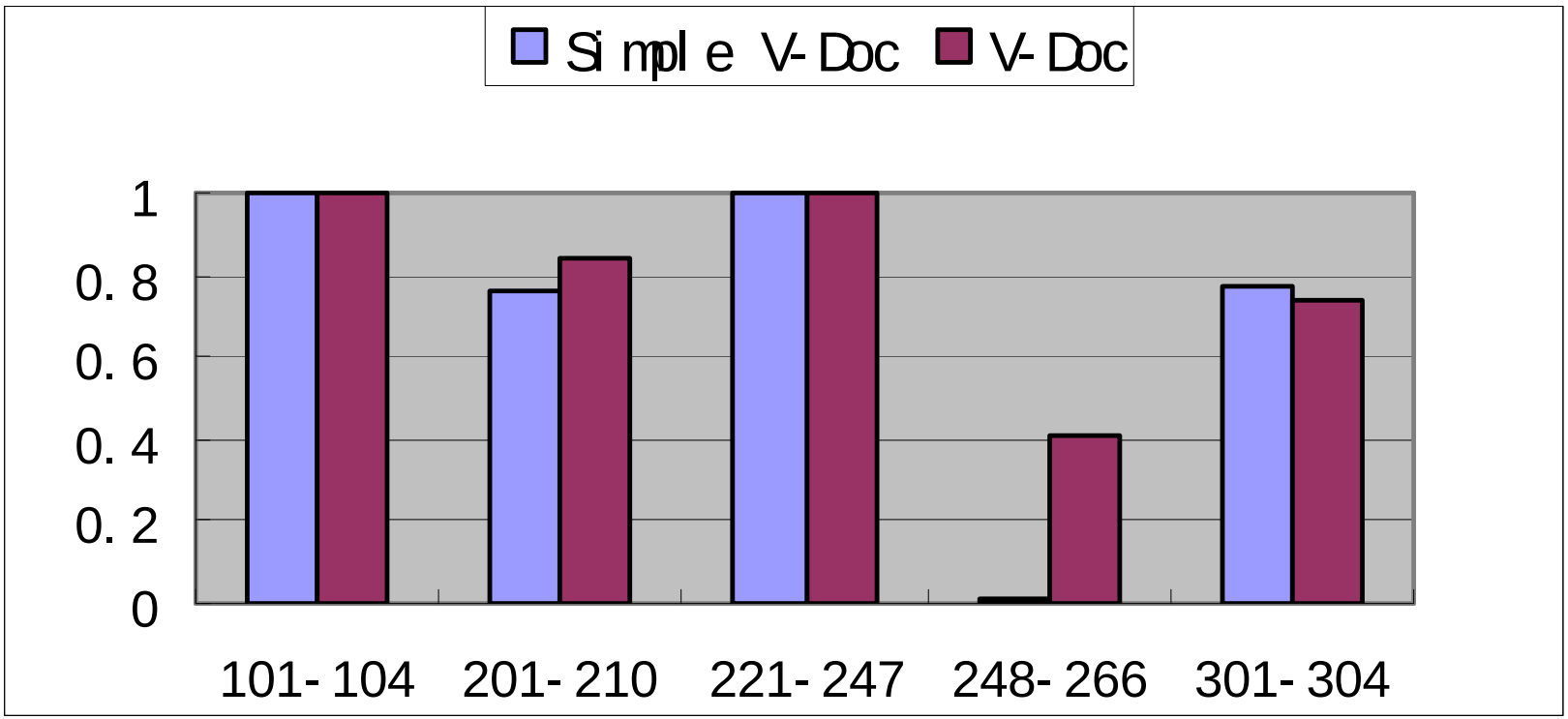
- Experiment on the OAEI 2005 benchmark tests
  - Test 101-104: No heterogeneity in linguistic feature
  - Test 201-210: Heterogeneity in linguistic feature
  - Test 221-247: Heterogeneity in structure
  - Test 248-266: The most difficult ones (heterogeneity)
  - Test 301-304: ontologies of bibliographic references
- Commodity PC
  - Intel Pentium 4, 2.4 GHz processor, 512M memory
  - Windows XP

Experiments — Setting(2)

- Parameters in constructing VD
  - Weighting local name, label and comment: 1.0, 0.5, 0.25
  - Damping factor along with blank node chain: 0.5
  - Weighting subject/predicate/object neighboring: 0.1
- Cosine (tfidf) is used to compute the similarity
- No cutoff in mapping selection, i.e. threshold=0
- Evaluation metrics: F-Measure

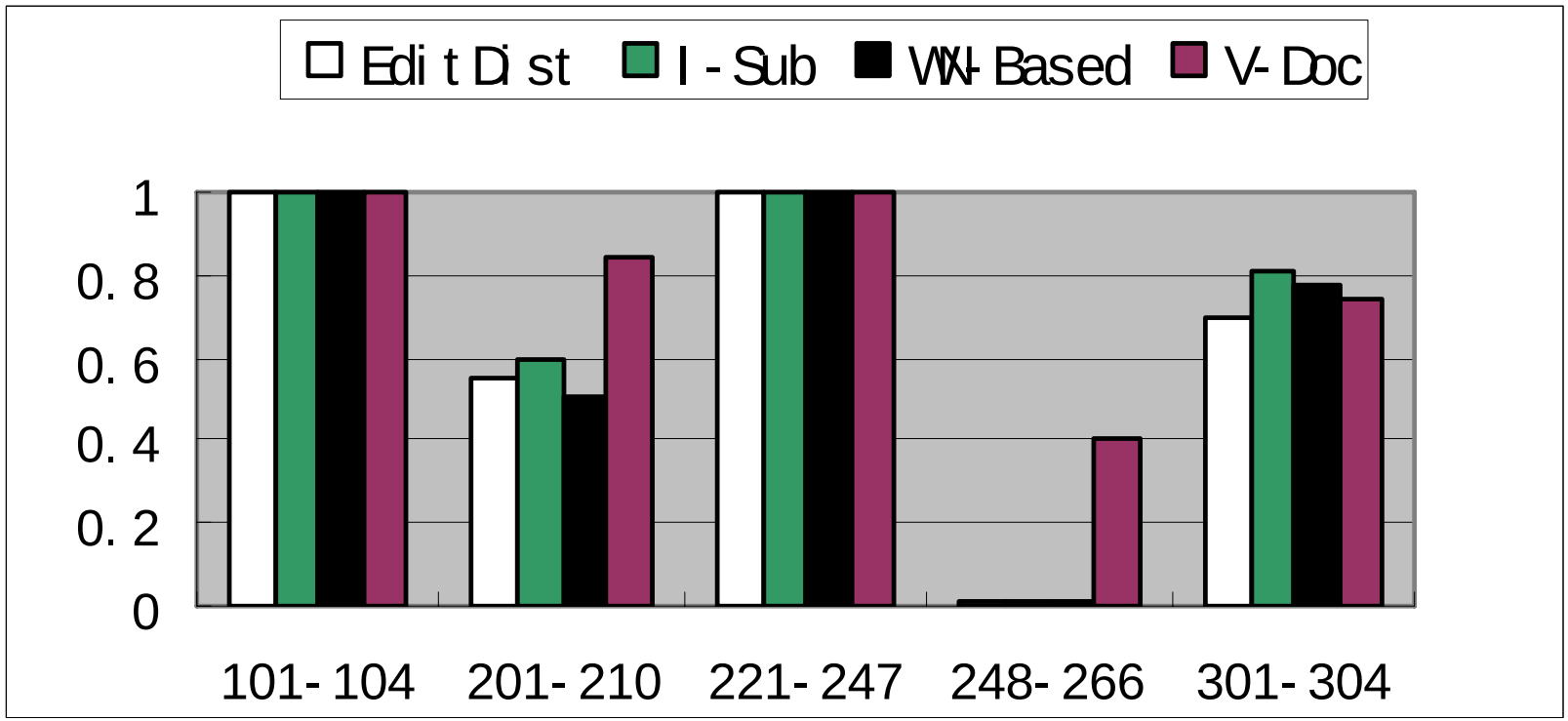
Experiments — Result (1)

- V-Doc VS Simple V-DOC (without neighboring infor)



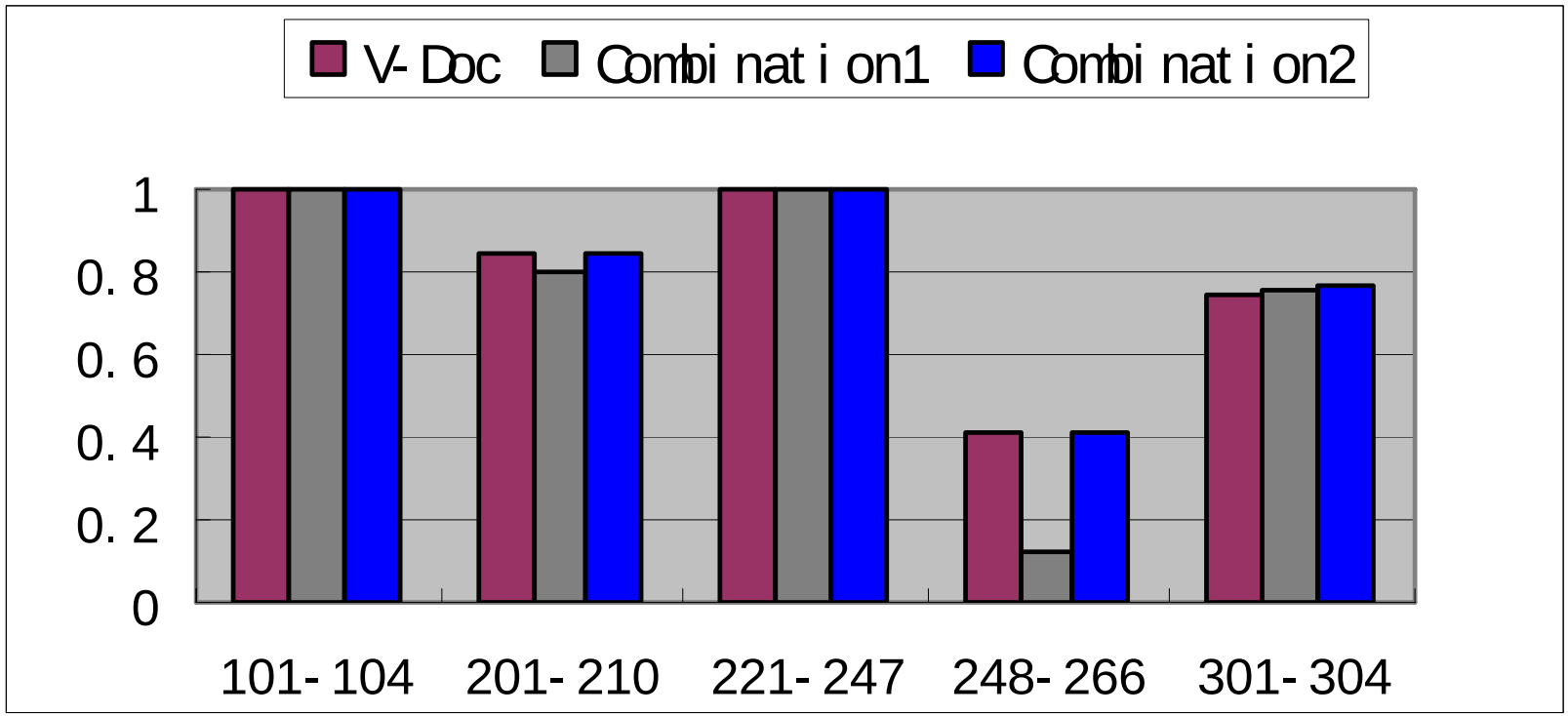
Experiments — Result (2)

- V-Doc VS other linguistic matching approaches



Experiments — Result (3)

- Combine V-Doc with EditDist or I-Sub



Experiments — Overall Result

- With average runtime per test

	101-104	201-210	221-247	248-266	301-304	Overall Avg.	Avg. Time
EditDistance	1.0	0.55	1.0	0.01	0.70	0.60	0.94(s)
I-Sub	1.0	0.60	1.0	0.01	0.81	0.61	1.00(s)
WN-Based	1.0	0.51	1.0	0.01	0.78	0.59	282(s)
Simple V-Doc	1.0	0.76	1.0	0.01	0.77	0.64	4.3(s)
V-Doc	1.0	**0.84**	1.0	**0.41**	**0.74**	**0.77**	**8.2(s)**
Combination1	1.0	0.80	1.0	0.12	0.76	0.68	9.4(s)
Combination2	1.0	0.85	1.0	0.41	0.77	0.78	9.8(s)

Concluding Remarks

- Virtual document
  - Incorporates both local descriptions and neighboring information
  - Is comprehensive and well-founded (RDF)
- V-Doc is a “linguistic matching”, but slightly combines structural information
  - Simple, Practical and Cost-effective
  - A trade-off between efficiency and accuracy

Concluding Remarks

No Silver Bullet

Acknowledgement

Q&A

Falcon at XObjects Group

http://xobjects.seu.edu.cn/project/falcon ...