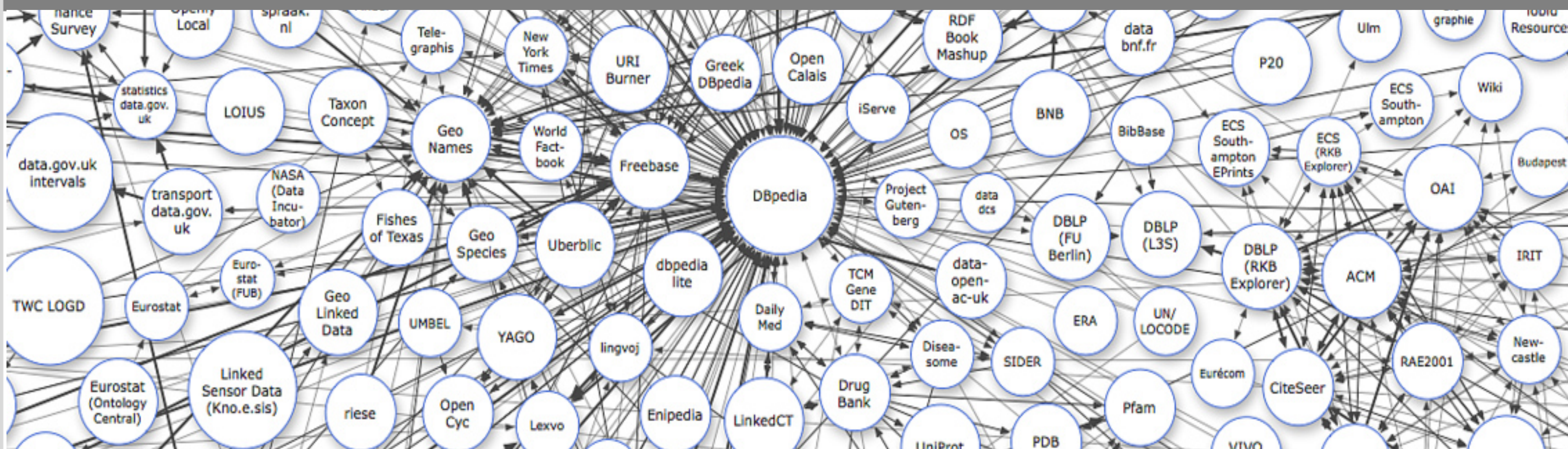
Efficient Graph-based Document Similarity

Christian Paul, Achim Rettinger, Aditya Mogadala, Craig A. Knoblock, Pedro Szekely

Institute of Applied Informatics and Formal Description Methods (AIFB)



Common task: Related-document Search

Query document

Apple breaks laptop sales record

Document Collection

⋮

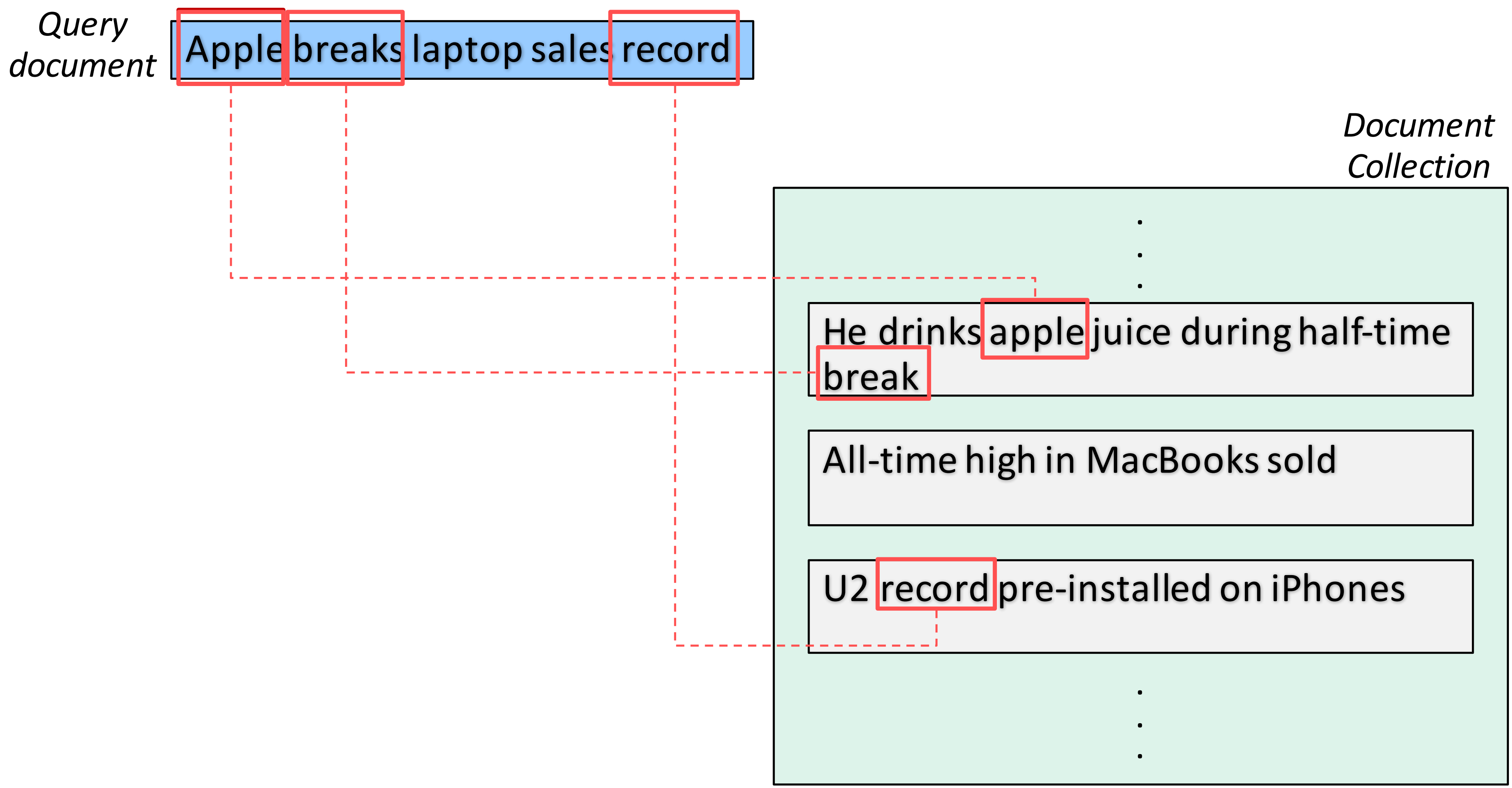
He drinks apple juice during half-time break

All-time high in MacBooks sold

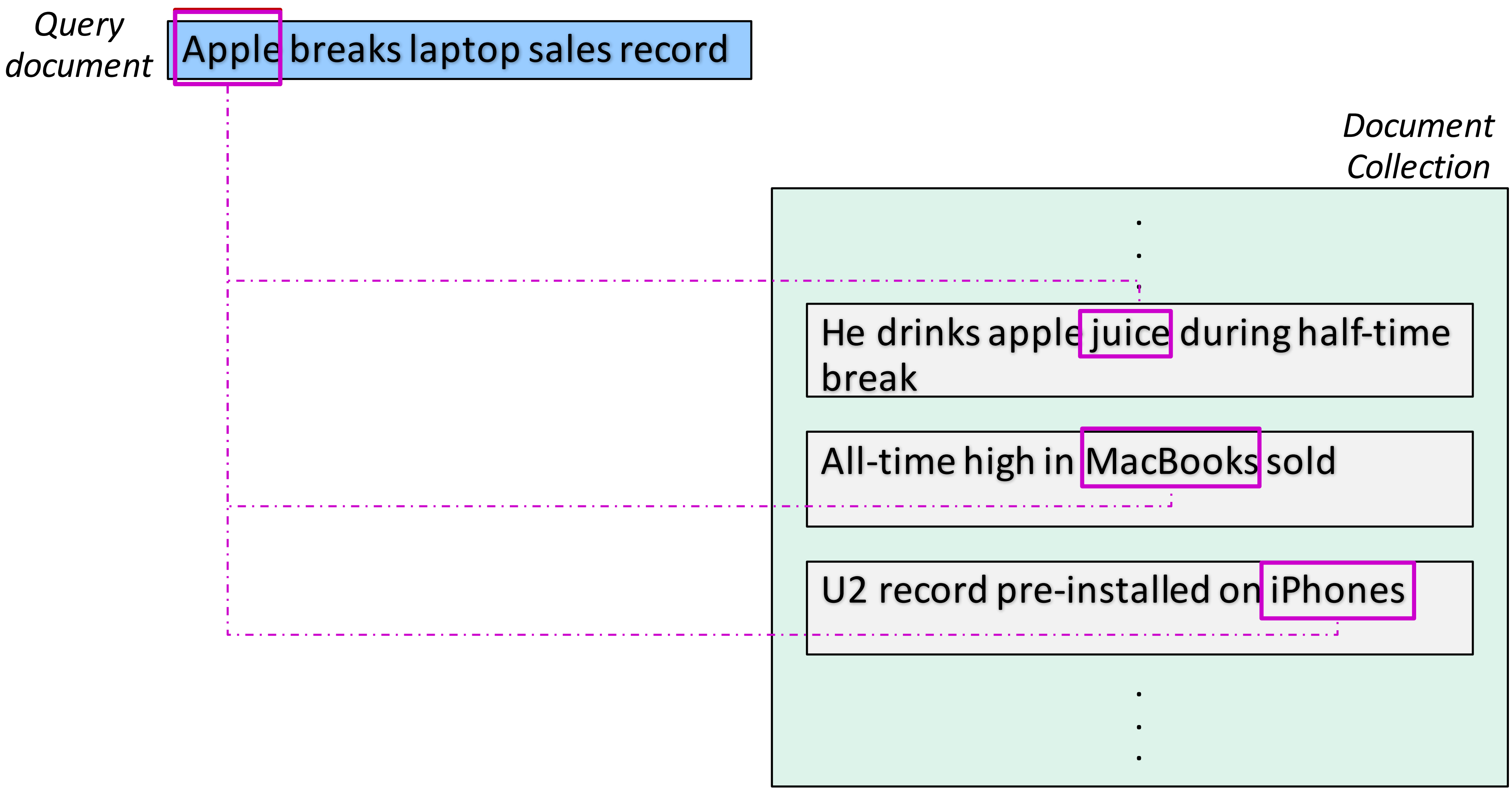
U2 record pre-installed on iPhones

⋮

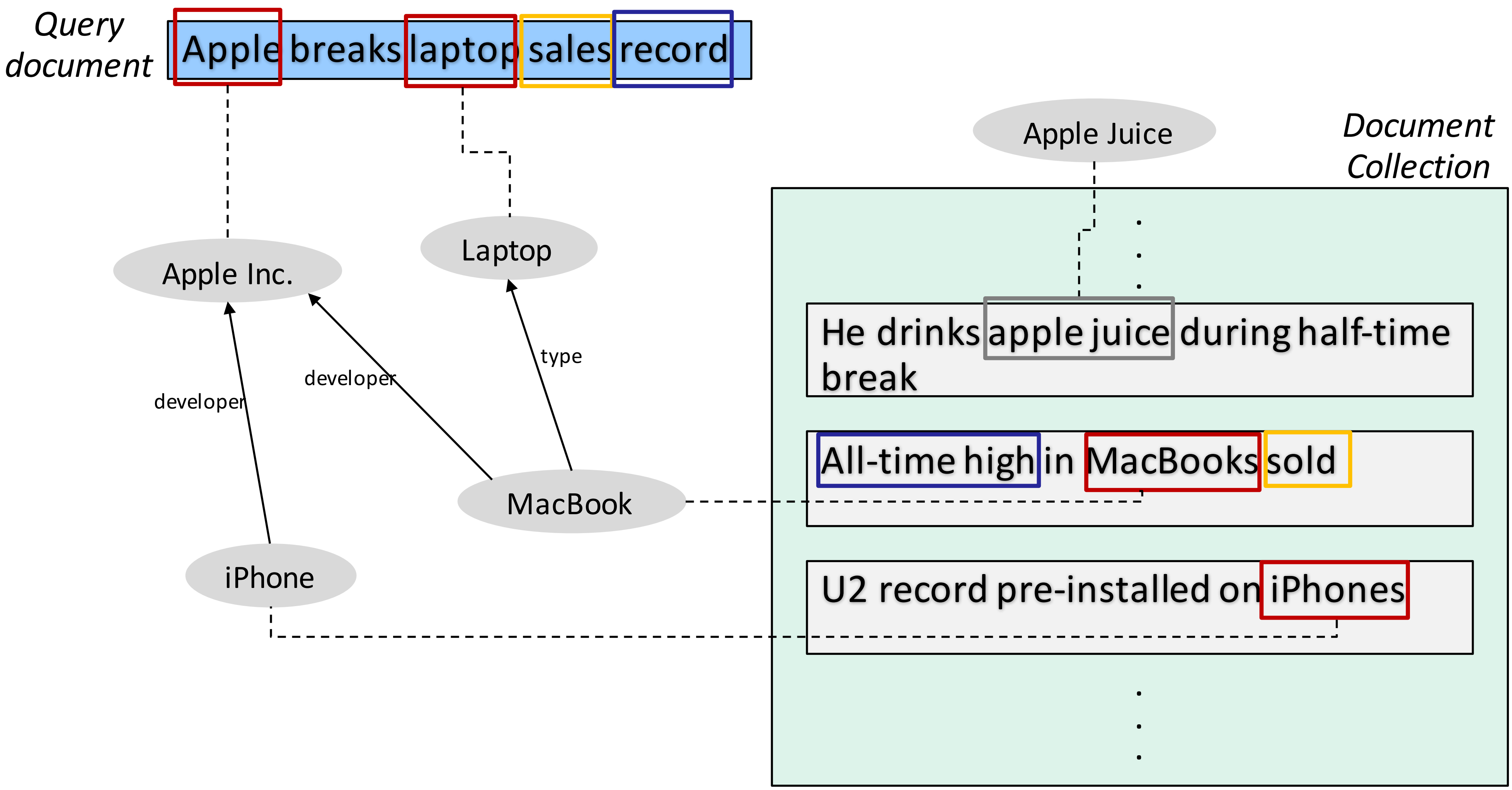
Matching words do not always indicate similarity



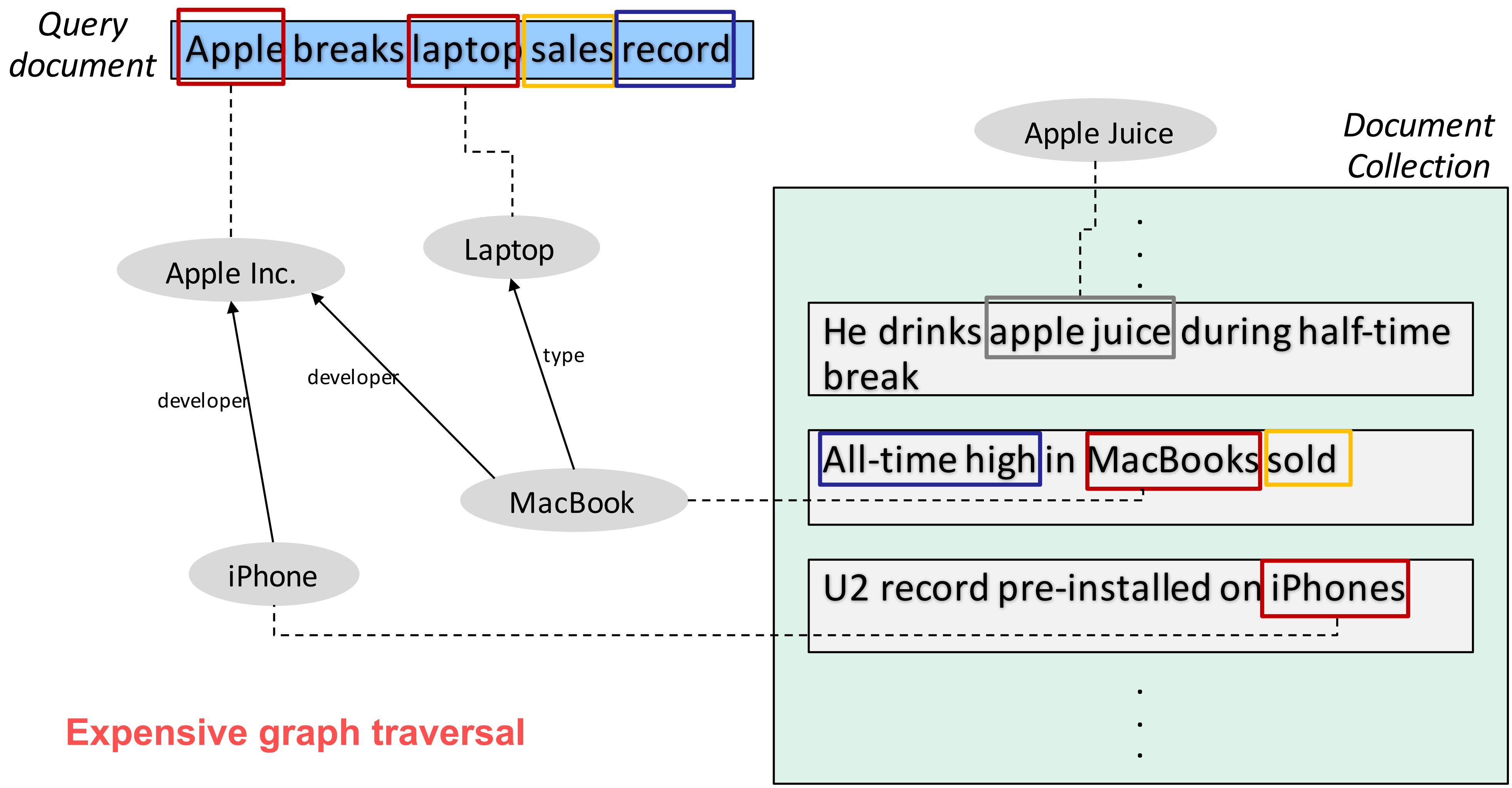
Word co-occurrence can be misleading, too



Semantic Technologies: resolve ambiguity & exploit relational knowledge



Semantic Technologies: resolve ambiguity & exploit relational knowledge



Expensive graph traversal

Related Work

TF-IDF,
Vector Space Model

Explicit Semantic Analysis (ESA) [GM07]

Salient Semantic Analysis (SSA) [HM11]

Distributional:
+ scalable, fast
- No explicit disambiguation and conceptual relations

Knowledge-based:
+ rich semantic knowledge
- expensive graph traversal

PathSim [SHY +11]
HeteSim [SKH +14]

Nunes et al.: Transversal doc. similarity [NKF +13]

Schuhmacher, Ponzetto: Graph Edit Distance [SP14]

AnnSim: 1-1 matching, hierarchical similarity [PVH +13]

Bridging the gap

Karlsruher Institut für Technologie

TF-IDF,
Vector Space Model

Explicit Semantic Analysis (ESA) [GM07]

Salient Semantic Analysis (SSA) [HM11]

Distributional:
- ***+ scalable, fast***
- ***- No explicit disambiguation and conceptual relations***

Efficient Graph-based Document Similarity

Knowledge-based:
- ***+ rich semantic knowledge***
- ***- expensive graph traversal***

PathSim [SHY +11]
HeteSim [SKH +14]

Nunes et al.: Transversal doc. similarity [NKF +13]

Schuhmacher, Ponzetto: Graph Edit Distance [SP14]

AnnSim: 1-1 matching, hierarchical similarity [PVH +13]

Core Contributions

- Scalable related-document search process
  - Graph traversal during pre-processing
  - Light-weight tasks at search time

We achieve similar computational efficiency as statistical approaches

Core Contributions

- Scalable related-document search process
  - Graph traversal during pre-processing
  - Light-weight tasks at search time

We achieve similar computational efficiency as statistical approaches

- Bag-of-entities document model & similarity
  - Document similarity as combination of pairwise entity similarities
  - Exploits **hierarchical** & **transversal** knowledge graph relations

In our experiments, we achieve higher correlation with human notion of document similarity than the competition

Related-document Search using Graph-based Similarity

1) Semantic Document Expansion

- Enrich query document with relational knowledge

2) Inclusion in corpus

- Store & index *expanded* document

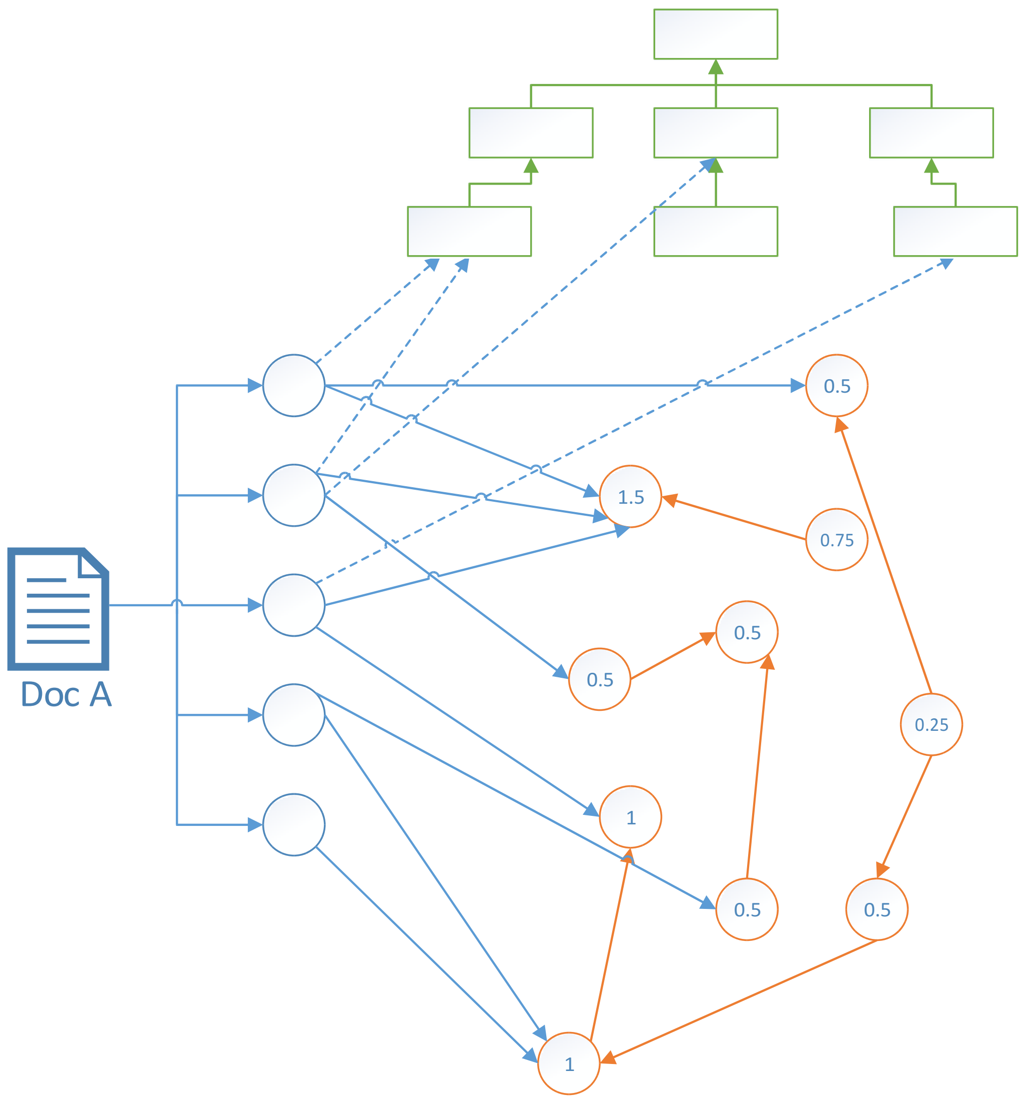
3) Pre-search

- Use inverted index to generate candidate set

4) Full search

- Entity-level, path-based similarities

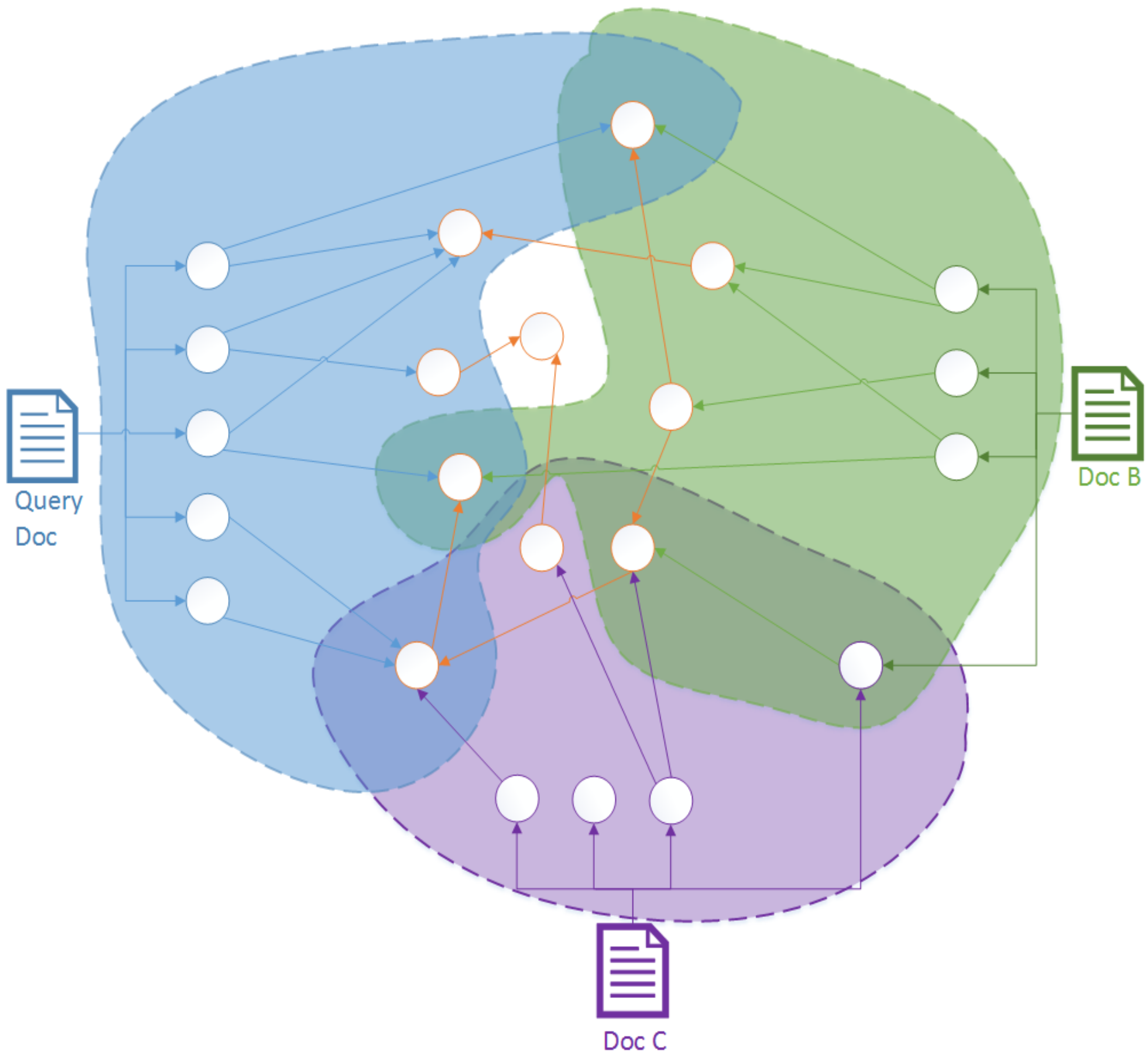
Semantic Document Expansion



- Enrich document annotations
- Hierarchically
  - Categories & their ancestors + hierarchical depths
- Transversally
  - Weight neighboring entities based on
    - number of paths
    - length of paths

$$w(e) = \sum_{l=1}^{L} \beta^{l} * \left| paths_{a,e}^{(l)} \right|$$

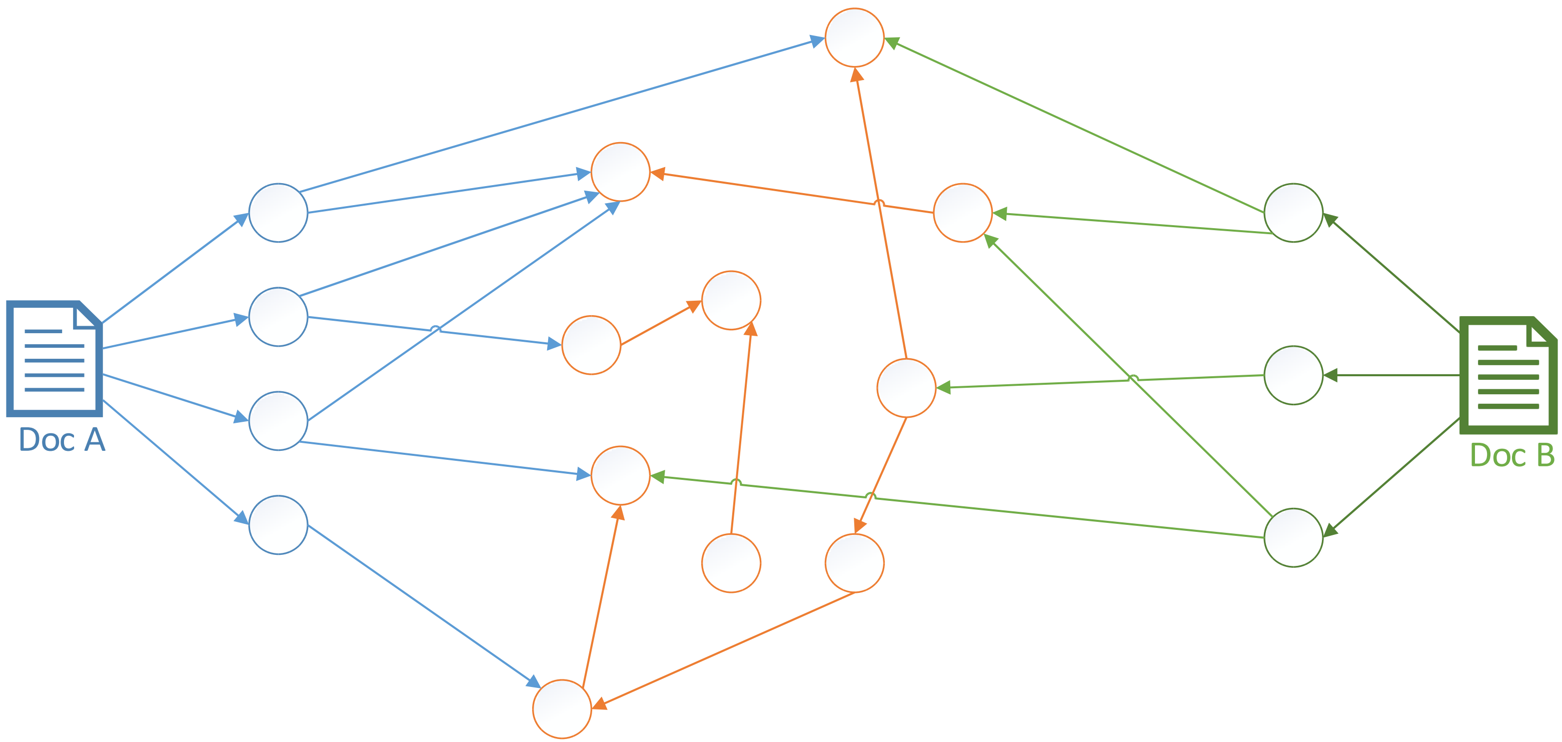
Pre-Search: Generate Candidate Set



- Inverted index from entities to documents
  - Retrieve candidates efficiently
- Assumption: Entity overlap → contextual similarity
  - Coarse, document-level assessment

Full Search: Graph-based Document Similarity

- For each candidate document, reconstruct query-candidate annotation subgraph - **hierarchical** & **transversal**

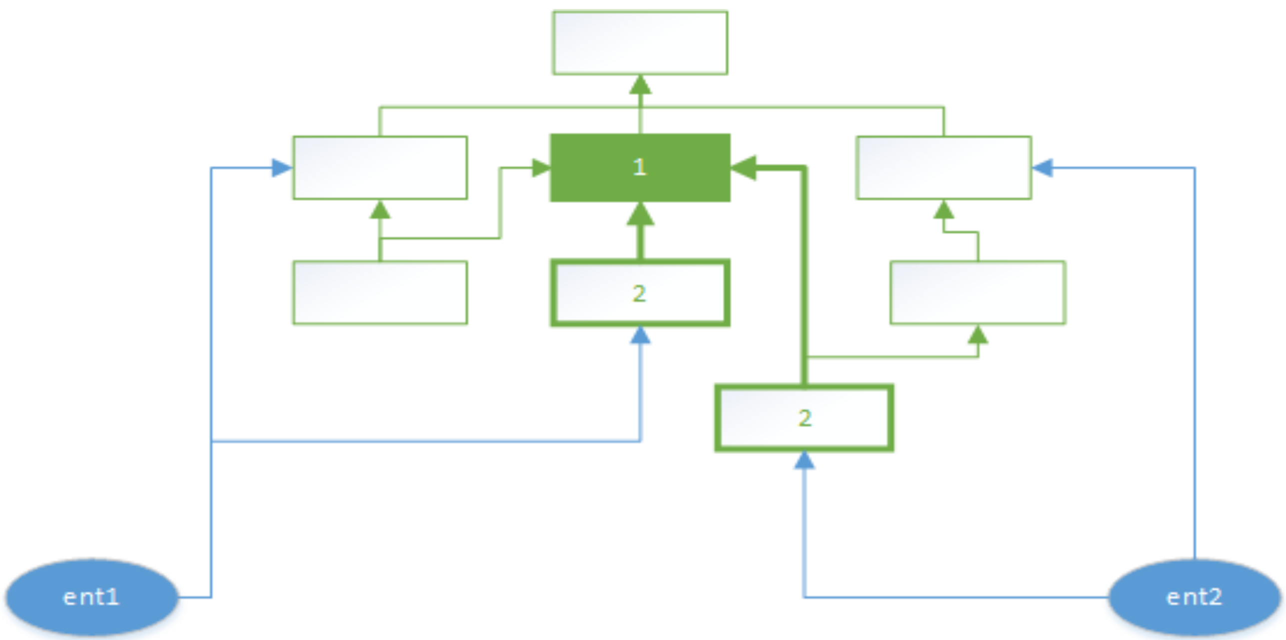


- ➢ Compute all pairwise entity similarity scores
- ➢ **Combine** into document score

Hierarchical entity similarity

- Using stored ancestors & depths to compute

$$hierSim_{dps}(x,y) = \frac{d(root, lca(x,y))}{d(root, lca(x,y)) + d(lca(x,y),x) + d(lca(x,y,),y)}$$

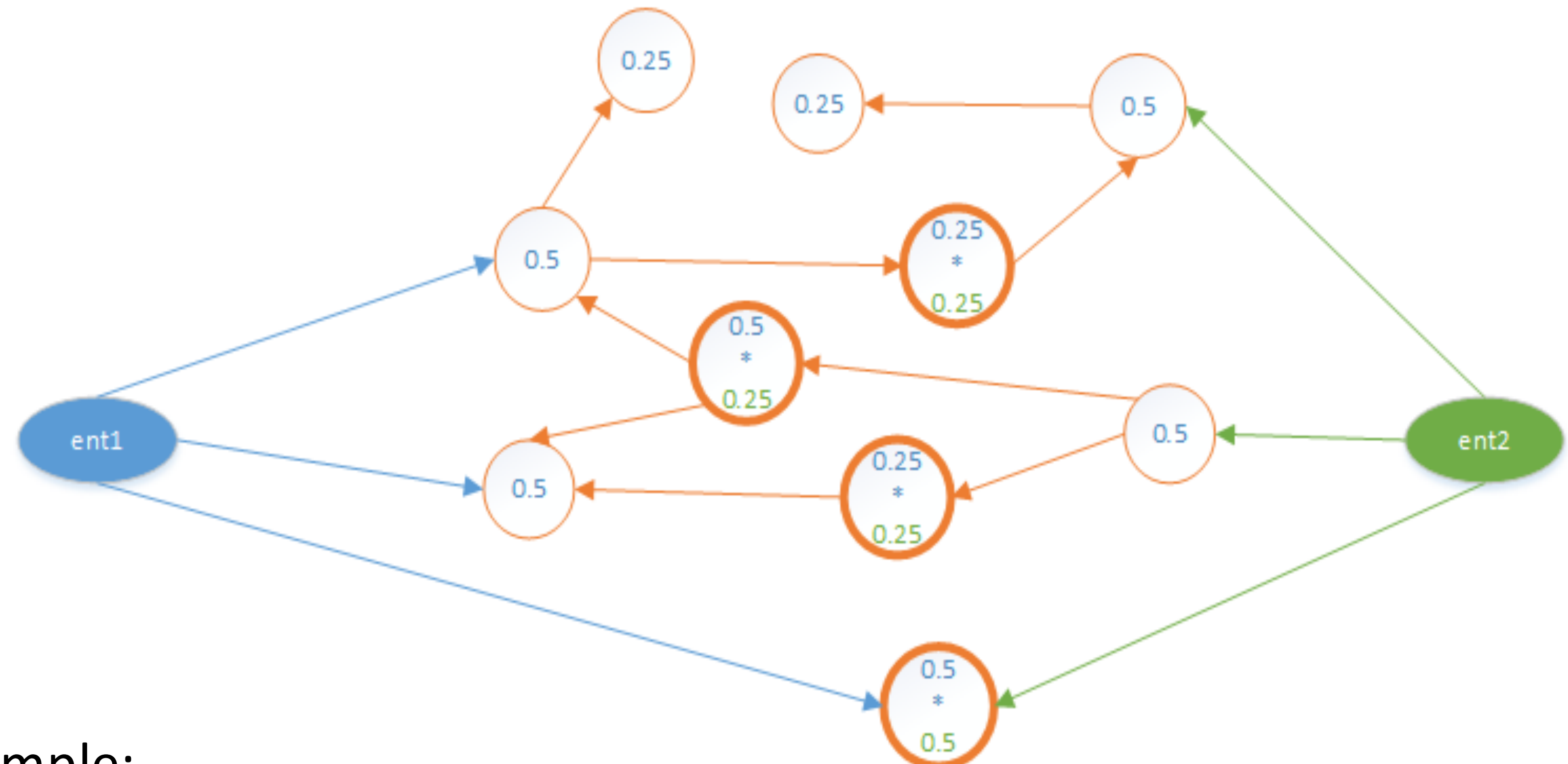


- Example: $hierSim_{dps}(ent1, ent2) = \frac{1}{1+2+2} = 0.2$

Transversal entity similarity

- Use stored neighbors & weights to compute:

$$transSim(a,b) = \sum_{l=1}^{L*2} \beta^{l} * \left| paths_{a,b}^{(l)} \right|$$



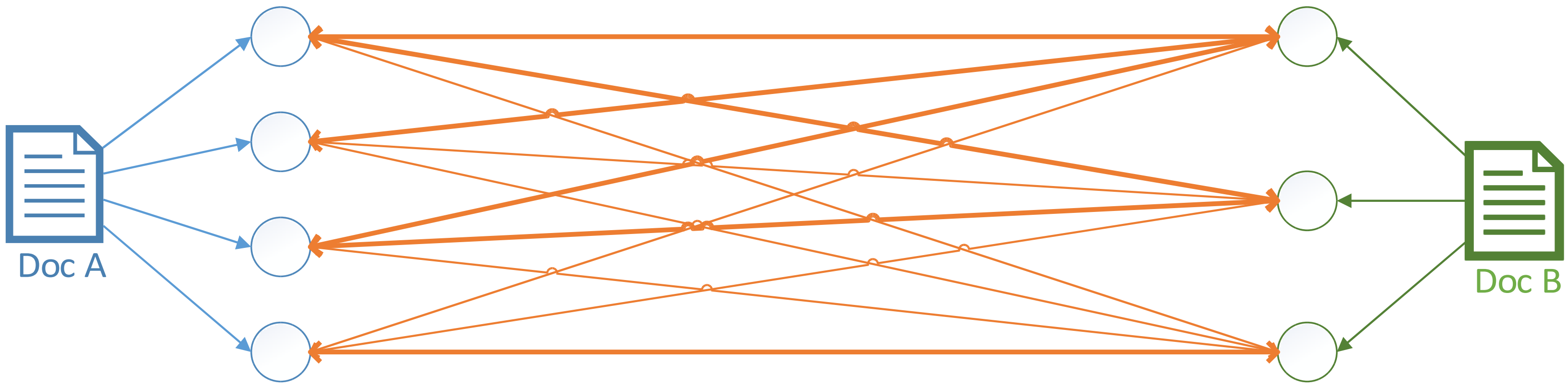
- Example:

$$transSim(ent1, ent2) = 0.5^2 + 2 * 0.25^2 + 0.5 * 0.25 = 0.5$$

Document similarity: bipartite graph of entity similarities

1. Annotation pair similarity: Combine transversal & hierarchical scores

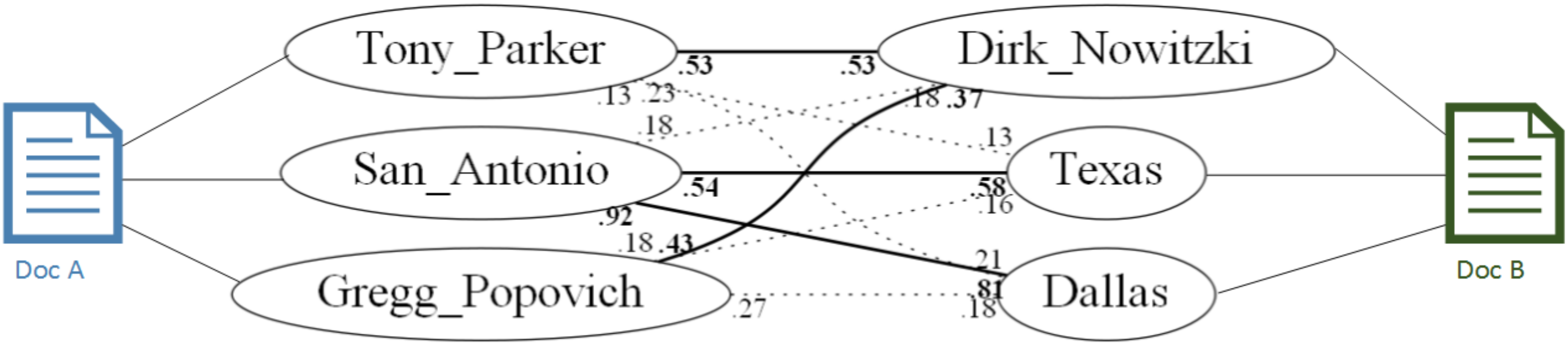
2. Determine *maxGraph:* for each annotation, choose max. score edge (bold)



3. Compute document score based on max. edges $(a_{1i}, matched(a_{1i}))$ for each annotation a_{1i} of Doc A:

$$docSim(docA, docB) = \frac{\sum_{a_{1i} \in A_1} (entSim_{ent}(a_{1i}, matched(a_{1i})))}{|A_1| + |A_2|}$$

Document similarity: DBpedia example



- Example documents score:

$$docSim(docA, docB) = \frac{0.53 + 0.92 + 0.43 + 0.53 + 0.58 + 0.81}{3 + 3} \approx 0.63$$

Evaluation

- Task: *Measure correlation with human notion of similarity*
- Datasets
  - **Document similarity**: Lee50[1]
  - **Sentence similarity**: 2012-MSRvid-Test[2], 2015-Images[3]
- ... using DBpedia and **X-LiSA**[ZR14] entity extractor

[1] https://webfiles.uci.edu/mdlee/LeePincombeWelsh.zip
[2] http://research.microsoft.com/en-us/downloads/38cf15fd-b8df-477e-a4e4-a4680caa75af/
[3] http://ixa2.si.ehu.es/stswiki/index.php/

Document Similarity: Lee50 corpus

- 50 short news articles (51 to 126 words)
- Gold standard set of full pairwise document similarity scores

➢ Outperforming baselines & competition:

- Statistical *(LSA, ESA, SSA)*
- Knowledge-based *(GED)*

		Correlation		
		r	ρ	μ
Baseline	$TF-IDF$	0.398	0.224	0.286
	$AnnOv$	0.59	0.46	0.517
Related	LSA	0.696	0.463	0.556
	SSA	0.684	0.488	0.569
	GED	0.63	-	-
	ESA	0.656	0.510	0.574
Ours	$\mathbf{GBSS}_{r=2}$	**0.712**	0.513	0.596
	$\mathbf{GBSS}_{r=3}$	0.704	**0.519**	**0.598**

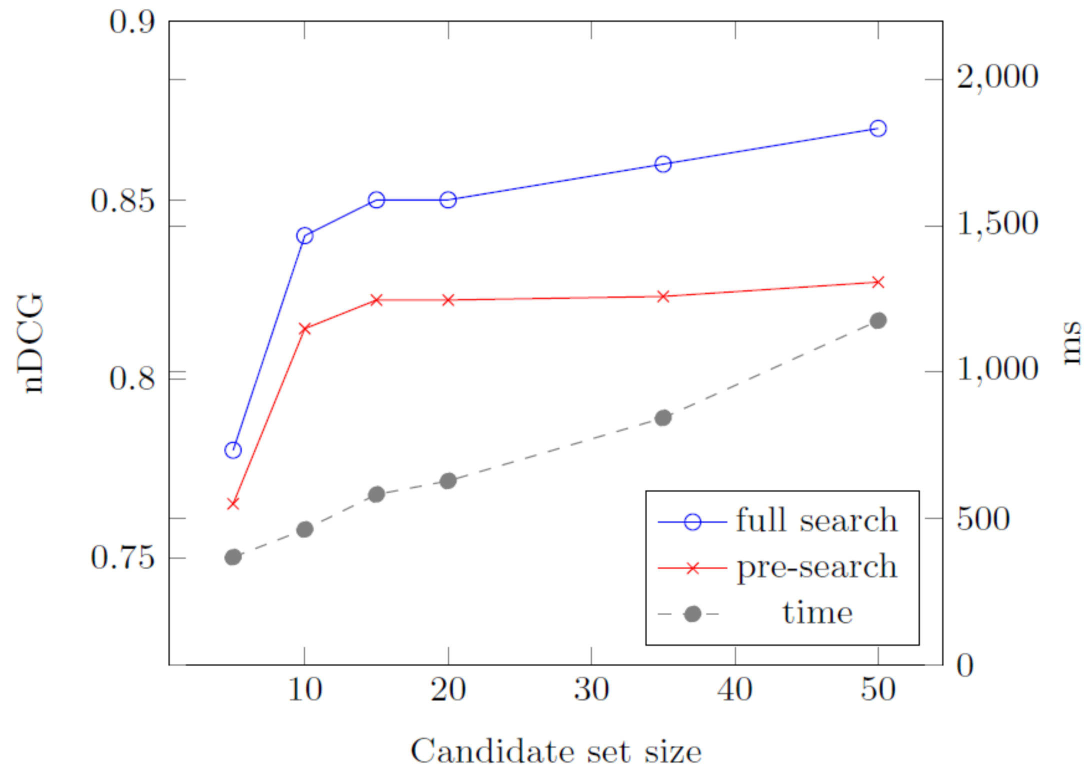
Sentence Similarity

- Compared to related unsupervised approaches
  (on texts with one or more extracted entities)
  - 2012-MSRvid-Test: Video descriptions from MSR Video Paraphrase Corpus
  - 2015-Images: Flickr image descriptions

- Outperforming baselines & competition
  - Statistical (Polyglot)
  - Knowledge-based (Tiantianzhu7, IRIT, WSL)

		Sentence Semantic Similarity	
		2012-MSRvid-Test	2015-Images
Baseline	STS-12	0.299	-
	STS-15	-	0.603
Related	Polyglot [3]	0.052	0.194
	Tiantianzhu7 [24]	0.594	-
	IRIT [6]	0.672	-
	WSL [22]	-	0.640
Ours	$\mathbf{GBSS}_{r=2}$	0.666	**0.707**
	$\mathbf{GBSS}_{r=3}$	**0.673**	0.665

Related-document Search: Pre-Search, Full Search & Efficiency



- Ranking score (nDCG) improves from Pre-Search to Full Search
- **Computation time** grows linearly with candidate set size
- Here: candidate set of size ~15 achieves high performance

Conclusion & Outlook

- Efficient Graph-based Document Similarity
  - … combines **hierarchical & transversal** relational knowledge
  - … **outperforms** related distributional & knowledge-based approaches, on both articles and sentences
  - … is computationally **efficient**: related-document search
- Lessons learned
  - Value of DBpedia for semantic similarity
  - The more entities (at least one) per document, the better:
    - Few entities: disambiguation helps
    - Many entities: *maxGraph* entity pairing emphasizes meaningful relations
- Resources (code, data, documents): *http://people.aifb.kit.edu/amo/eswc2016/*

References I

- **[TMS08]** Thiagarajan, Manjunath, Stumptner. Computing semantic similarity using ontologies. In *ISWC 08, the International Semantic Web Conference (ISWC)*, 2008.
- **[LD08]** Lemaire, Denhière. Effects of high-order co-occurrences on word semantic similarities.
- **[GM07]** Gabrilovich, Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, pages 1606–1611, 2007.
- **[HM11]** Hassan, Mihalcea. Semantic relatedness using salient semantic analysis. In *AAAI*, 2011.
- **[SP14]** Schuhmacher, Ponzetto. Knowledge-based graph document modeling. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM ’14.
- **[NKF+13]** Nunes, Kawase, Fetahu, Dietze, Casanova, Maynard. Interlinking documents based on semantic graphs. *Procedia Computer Science*, 22:231–240, 2013.
- **[PSA08]** Potthast, Stein, Anderka. A wikipedia-based multilingual retrieval model. In *Advances in Information Retrieval*, pages 522–530. Springer, 2008.
- **[SHY+11]** Sun, Han, Yan, Yu, Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *VLDB’11*, 2011.
- **[SKH+14]** Chuan, Xiangnan, Yue, Yu, Bin. Hetesim:A general framework for relevance measure in heterogeneous networks. *IEEE Transactions on Knowledge & Data Engineering*.
- **[PVH+13]** Palma, Vidal, Haag, Raschid, Thor. Measuring relatedness between scientific entities in annotation datasets. In *Proceedings of the International Conference on Bioinformatics,Computational Biology and Biomedical Informatics*, BCB’13.
- **[ZR14]** Zhang, Rettinger. X-lisa: Cross-lingual semantic annotation. *Proceedings of the VLDB Endowment (PVLDB), the 40th International Conference on Very Large Data Bases (VLDB).*
- **[KJC+15]** Pavan Kapanipathi, Prateek Jain, Chitra Venkataramani, Amit Sheth. Hierarchical interest graph, 21 January 2015. wiki.knoesis.org/index.php/Hierarchical_Interest_Graph, last accessed 07/15/2015

References II

- **[LIJ+15]** Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., et al.: Dbpedia-a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web 6(2), 167-195 (2015)