A Novel Approach for Patent Similarity Measurement Based on Sequence Alignment

Presenter: Jinghong Li

CONTENTS

1 Introduction

Introduction

- Significance
  - According to many surveys of authorities, patents cover more than 90% latest technical information of the world, of which 80% would not be published in other forms (Zha & Chen 2010). Discovering the technical intelligence via patents analysis is increasingly vital.
  - Patent similarity measurement is one of fundamental building blocks for patent analysis, since it is able to derive technical intelligence efficiently, but also can detect the risk of infringement and evaluate whether the invention meets the criteria of novelty and innovation.

- Research status of patent similarity measurement
  - the bibliographic information based approaches
  - the lexical based approaches

The bibliographic information based approaches

IPC classification	
Type	**Contents**
A	Human Necessities
B	Performing Operations; Transporting
C	Chemistry; Metallurgy
D	Textiles; Paper
E	Fixed Constructions
F	Mechanical Engineering; Lighting; Heating; Weapons; Blasting Engines or Pumps
G	Physics
H	Electricity

IPC System

- hierarchical classification system
- suitable for measuring patent similarity.

Drawbacks:

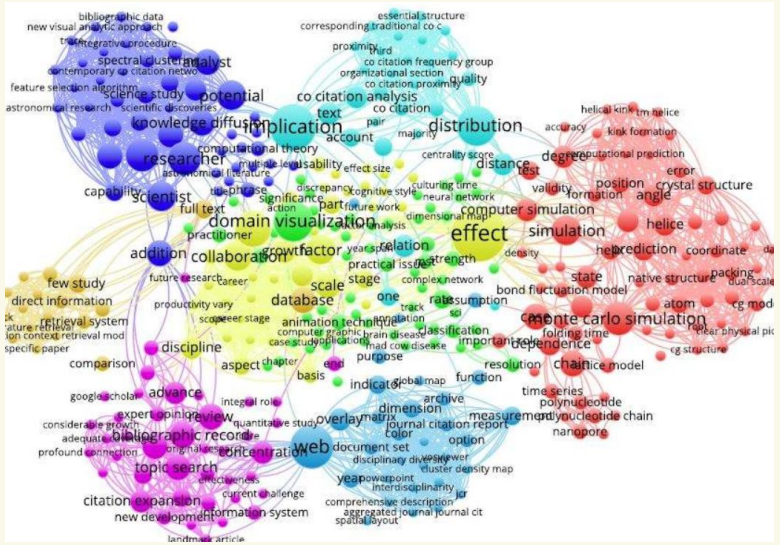
- heavily dependent on existing technologies
- uncertainty

Citation Network

- serve as a proxy to assess similarity with bibliographic coupling network or co-citation network

Drawbacks:

- new patents tend to be less cited
- some patent databases do not provide citation information



The lexical based approaches

- **Keywords co-occurrence**
  - Two patents are similar with each other if they share a high degree of common keywords.

Drawbacks:

- heavily relies on the keyword choices and language style of the inventors
- insufficient to reflect specific technological key concepts and relations

- **Subject-Action-Object (SAO) semantic analysis**
  - stressing semantic similarity with the concept of *function* -- “the action changing a feature of any object”
  - describes a relation between components in the patent documents

Drawbacks:

- Functional relations
- Equal weight
- Ignore semantic direction and word order

Introduction

- Objectives – **an improved approach for patent similarity measurement**
  - functional and non-functional relations
  - semantic direction of each sequence structure
  - word order information of each component
  - different weight to each sequence structure

2 Related Works

Patent Similarity Measurement based on SAO structures

- Bergmann et al. (2008) and Park, Yoon & Kim (2012) utilized SAO based semantic technological similarities to evaluate the risk of patent infringement.
- Choi, Park, Kang, Lee & Kim (2012) categorized the SAO structures extracted from patent documents to build a technology tree for technology planning with the help of similarity measurement method.
- The evolving technological trend for R&D planning was identified by Yoon & Kim (2011) by constructing a SAO semantic patent network based on the internal similarities between patents.
- Sternitzke & Bergmann (2009) focused on how to use SAO structures to improve the accuracy of comparison methods to evaluate patent similarities.

Drawbacks:

- assign the same weight to each SAO structure
- omit the word order information
- ignore non-functional relations

Patent Similarity Measurement based on SAO structures

- Wang et al. (2019) has constructed a DWSAO indicator through assigning different weights to SAO structures for measuring patent similarity.

Drawbacks:

- It neglects the influence of the number of SAO structures of patents, which may result in the phenomenon that patents with high similarity values are actually not similar.
- It is not a symmetrical indicator because of the weighting strategy.
- omit the word order information
- ignore non-functional relations

WordNet for Semantic Similarity of Words

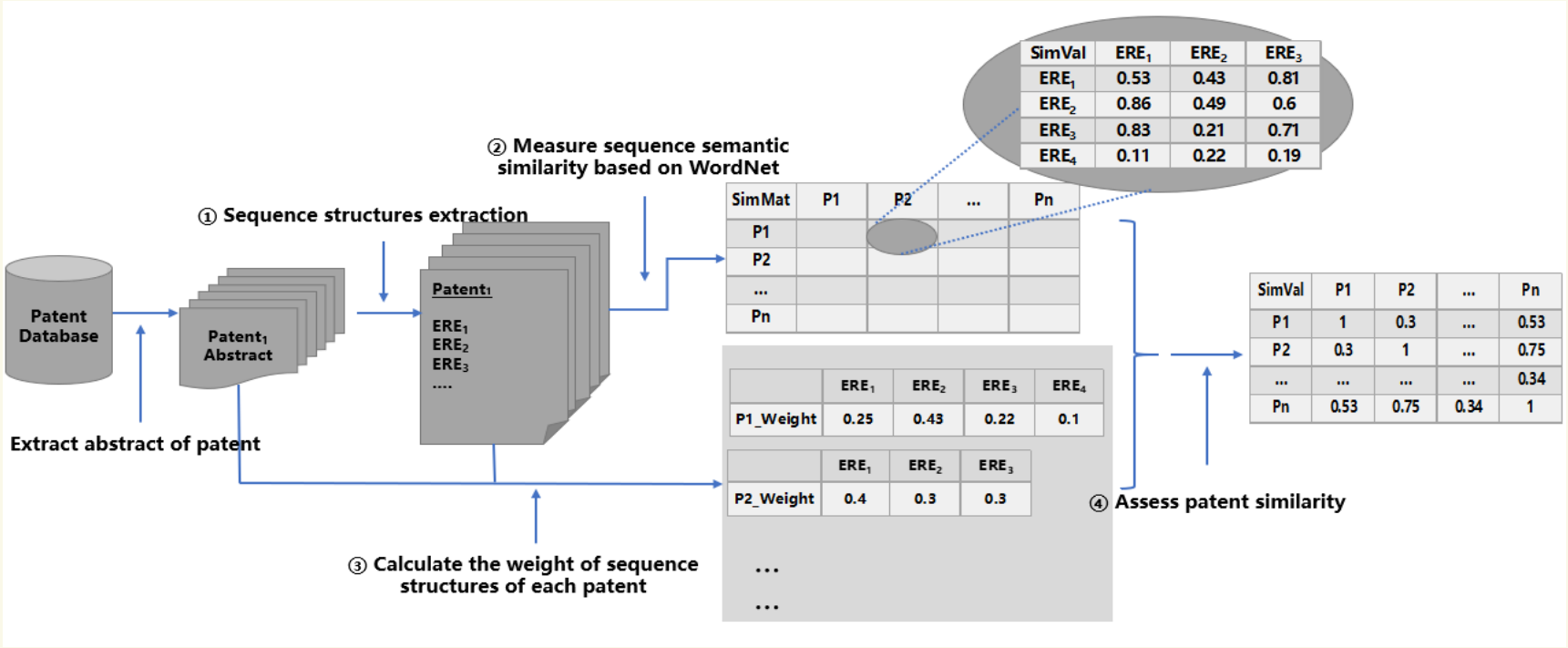
- WordNet is a lexical database which groups English concepts into sets of synonyms called “synsets” and constructs the hierarchical structure to connect “synsets” by means of hypernym/hyponym relations. Just because of this property, WordNet is commonly used to calculate the semantic similarity of concepts.
- The IC-based approach is utilized in this paper, which can be formally defined as follows (Lin 1998):

$$sim(c_1, c_2) = \frac{2 \times IC(LCS)}{IC(c_1) + IC(c_2)}$$

- Note that a word may express different meaning (concept) in different context, viz. polysemy. This paper uses the concepts corresponding to the highest similarity between two words. The similarity of two words can be defined as follows:

$$sim(word_1, word_2) = \max_{c_i \in Syn_1} \max_{c_j \in Syn_2} sim(c_i, c_j)$$

3 Methodology



The procedure for measuring patent similarity

- **Sequence structures extraction**
- **Similarity between sequence structures**
- **Weight estimation of sequence structures of each patent**
- **Patent similarity assessment**

3.1 Sequence structures extraction

- General NLP techniques and tools
  - Stanford CoreNLP (Manning et al. 2014)
  - OpenIE (Saha 2018)
  - …

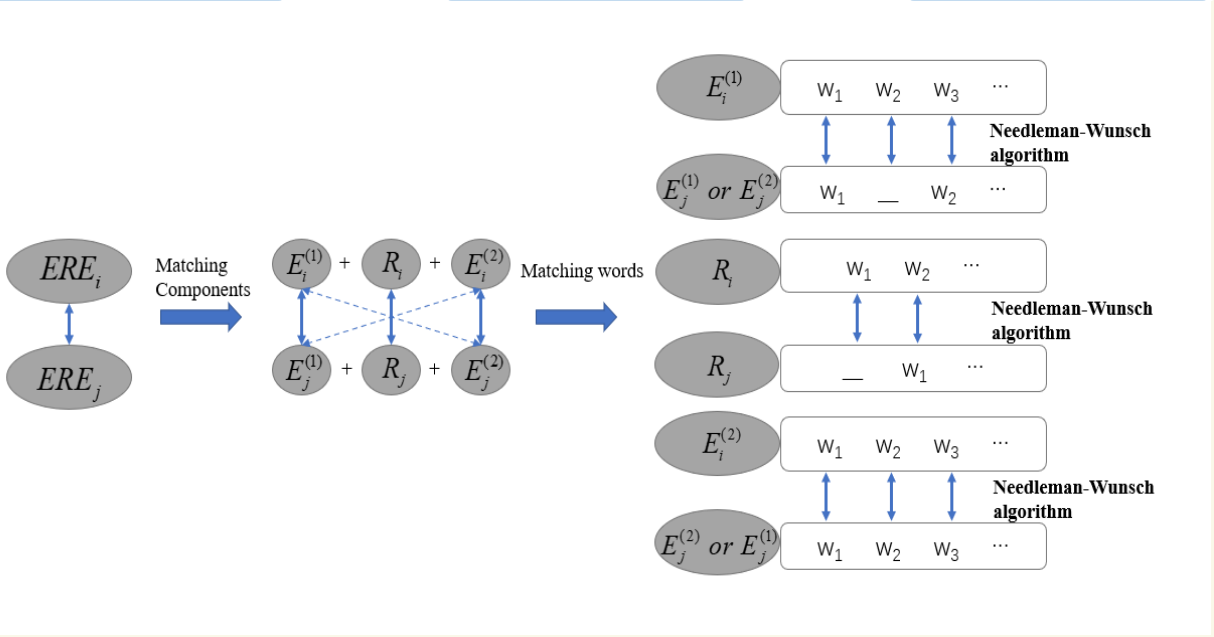
  Drawbacks: the performance and accuracy are not satisfactory in most cases, especially for extracting domain-specific entities and semantic relations.

- Chen et al. (2020) have proposed a promising patent information extraction framework. This framework is used here to extract the sequence structures mentioned in the patent documents.

Chen, L., Xu, S., Zhu, L., Zhang, J., Lei, X., & Yang, G. (2020). A deep learning based method for extracting semantic information from patent documents. DOI: 10.1007/s11192-020-03634-y

3.2 Similarity between sequence structures

Define semantic direction in accordance to relation type → Align the words by Needleman-Wunsch Algorithm → Calculate words' semantic similarity by WordNet



Each sequence structure consists of three components: $E^{(1)}$, R and $E^{(2)}$.

After extracting sequence structures, each patent can be represented by a collection of different number of sequence structures. Patent similarity calculation problem can be transformed to compute the similarity between the collections of sequence structures.

This subsection illustrates how to calculate the semantic similarity between two sequence structures:

- Define semantic direction to align the components from different structures
- Align the words in each component

3.2 Similarity between sequence structures

- We defined 4 types of semantic directions in accordance to the type of semantic relation.

	Relation Type	Semantic Direction
1	spatial relation	Undirected
2	part-of	$E^{(1)} \leftarrow E^{(2)}$
3	causative relation	$E^{(1)} \leftarrow E^{(2)}$
4	operation	$E^{(1)} \leftarrow E^{(2)}$
5	made-of	$E^{(1)} \leftarrow E^{(2)}$
6	instance-of	$E^{(1)} \rightarrow E^{(2)}$
7	attribution	$E^{(1)} \leftarrow E^{(2)}$
8	generate	$E^{(1)} \leftarrow E^{(2)}$
9	purpose	$E^{(1)} \leftarrow E^{(2)}$
10	in-manner-of	$E^{(1)} \leftarrow E^{(2)}$
11	alias	Bidirectional
12	formation	$E^{(1)} \rightarrow E^{(2)}$
13	comparison	Undirected
14	measurement	$E^{(1)} \leftarrow E^{(2)}$
15	others	Undirected

3.2 Similarity between sequence structures

- Align the components from different sequence structures
  - The sequence structures are both single-direction.

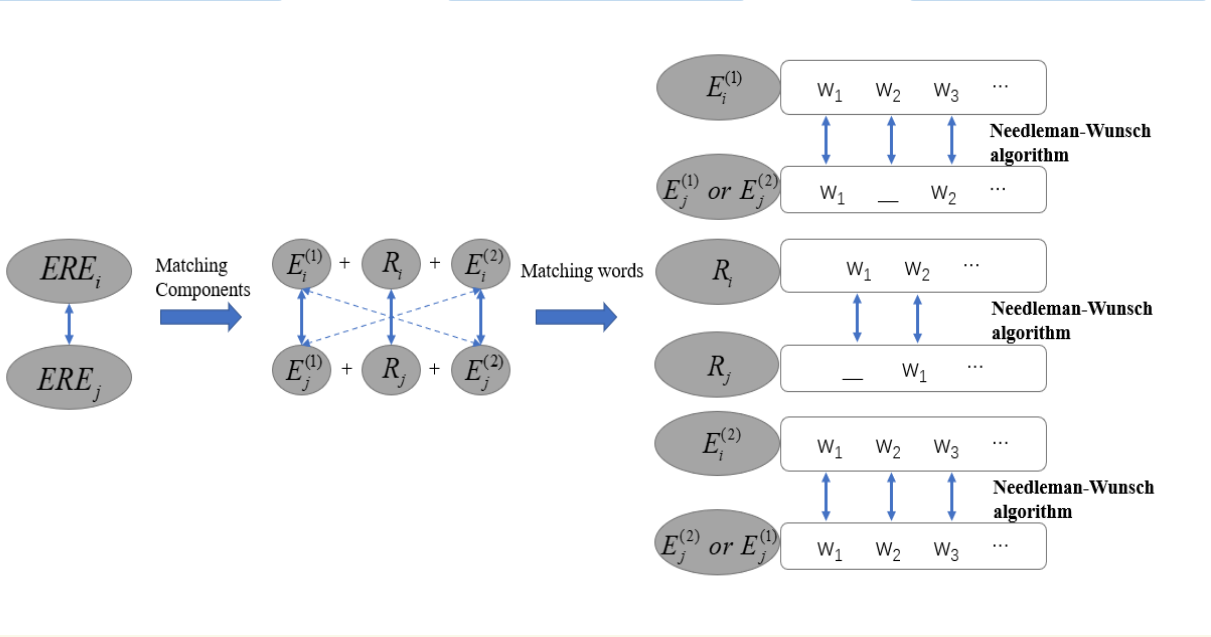
$$sim(ERE_i, ERE_j) = \begin{cases} \dfrac{sim(E_i^{(1)}, E_j^{(1)}) + sim(R_i, R_j) + sim(E_i^{(2)}, E_j^{(2)})}{3}, E_i^{(1)} \text{ matches } E_j^{(1)} \text{ and } E_i^{(2)} \text{ matches } E_j^{(2)} \\ \dfrac{sim(E_i^{(1)}, E_j^{(2)}) + sim(R_i, R_j) + sim(E_i^{(2)}, E_j^{(1)})}{3}, E_i^{(1)} \text{ matches } E_j^{(2)} \text{ and } E_i^{(2)} \text{ matches } E_j^{(1)} \end{cases}$$

  - Sometimes, it is very difficult to judge the semantic direction only from the component R (relations). Of course, there exist bidirectional relations.

$$sim(ERE_i, ERE_j) = max \begin{cases} \dfrac{sim\left(E_i^{(1)}, E_j^{(1)}\right) + sim(R_i, R_j) + sim\left(E_i^{(2)}, E_j^{(2)}\right)}{3} \\ \dfrac{sim\left(E_i^{(1)}, E_j^{(2)}\right) + sim(R_i, R_j) + sim\left(E_i^{(2)}, E_j^{(1)}\right)}{3} \end{cases}$$

3.2 Similarity between sequence structures

Define semantic direction in accordance to relation type → Align the words by Needleman-Wunsch Algorithm → Calculate words' semantic similarity by WordNet

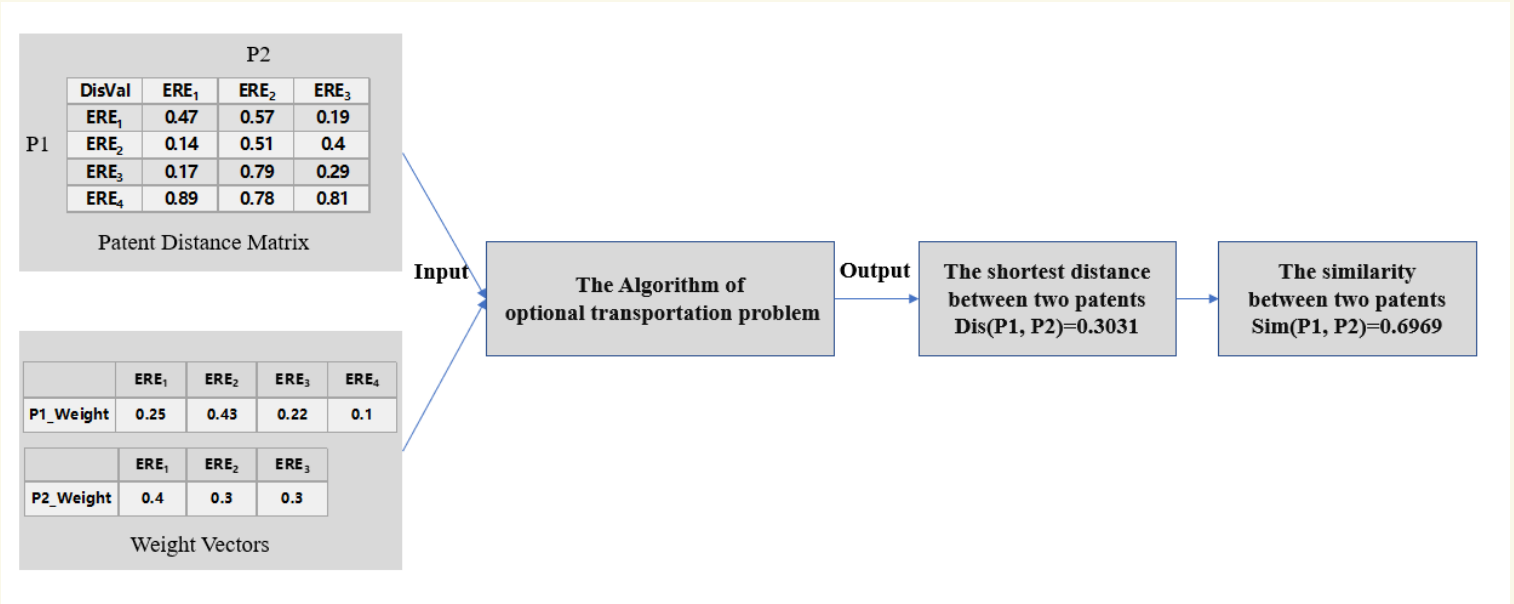


After that, we can get the patent similarity matrix between two patents.

Align the words in each component

- In most cases, these components are expressed with multiple words.
- Words order information should be considered.
- **Needleman-Wunsch algorithm** is utilized here to construct the correspondences of words.

3.4 Patent similarity assessment



The procedure of calculating the similarity between two patents

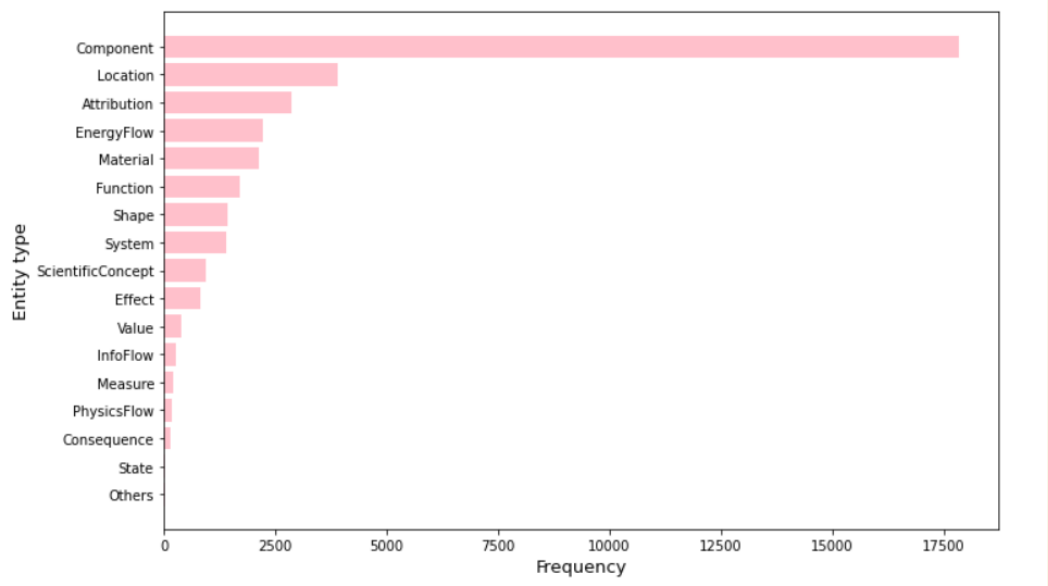
In order to make full use of all the information, patent similarity measurement problem can be transformed into the well-known optimal transportation problem (Xu et al. 2019).

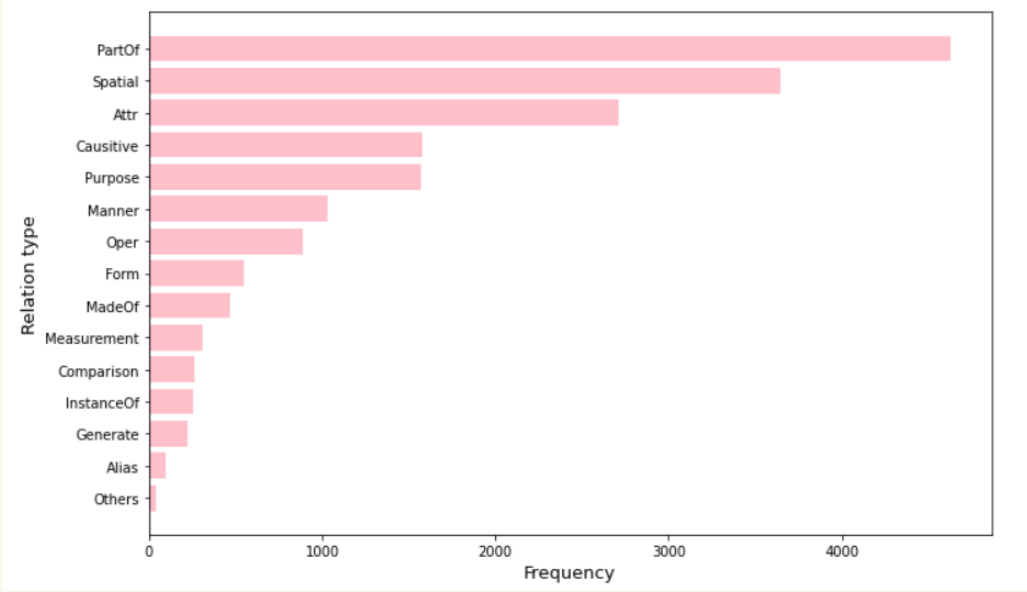
Xu S., Zhai D., Wang F., An X., Pang H., and Sun Y., 2019. A Novel Method for Topic Linkages between Scientific Publications and Patents. Journal of the Association for Information Science and Technology, Vol. 70, No. 9, pp. 1026-1042.

4 Case Study

4.1 Dataset

- originate from Chen et al. (2020) [1]
- *related to thin film head* subfield in the field of *hard disk drive*
- contains 1,010 patent documents and 18,264 sequence structures
- 17 entity types and 15 semantic relation types





[1] https://github.com/awesome-patent-mining/TFH_Annotated_Dataset

4.1 Dataset

- Note that, in this dataset, there are 84 pairs of patents coming from the same patent family. That is, they should have higher similarity than others.

- These patents can be used to assess the effectiveness and performance of our method. If a method can better identify these 84 pairs of patents, its performance should be better.

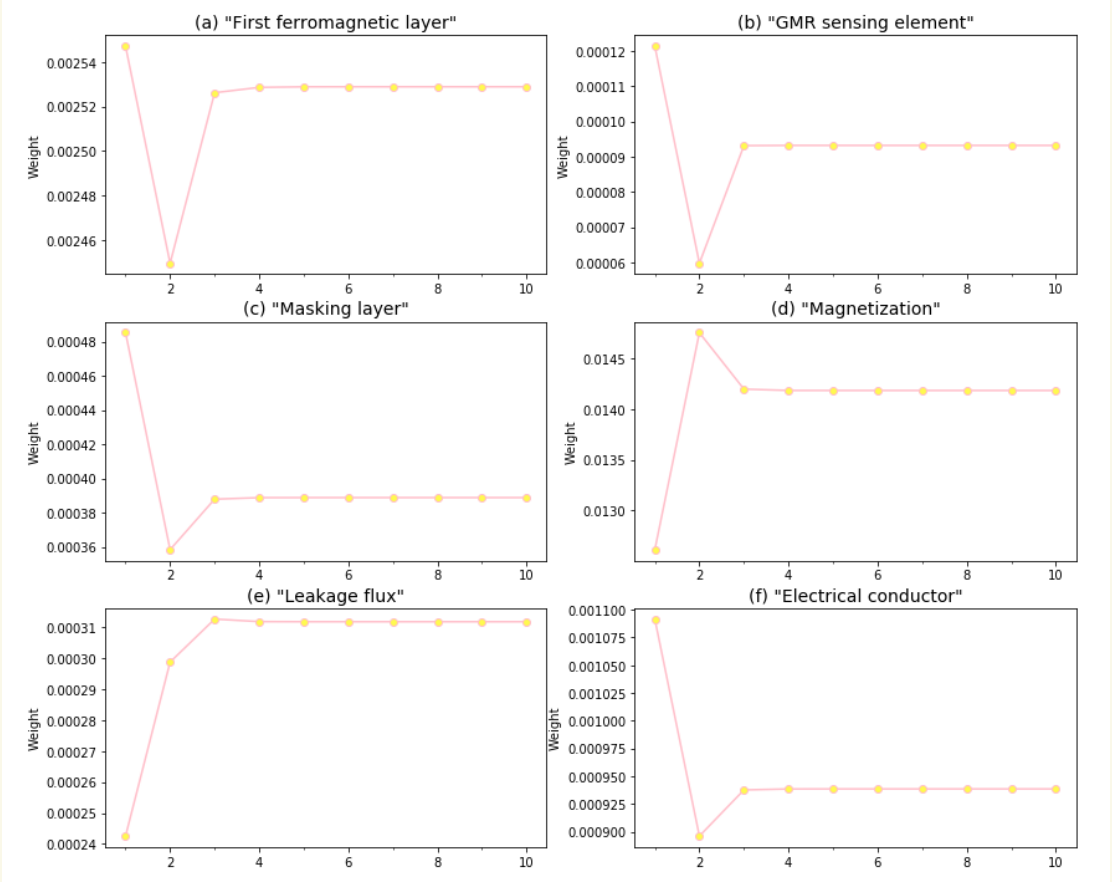
4.2 Experimental setup

- Semantic similarity calculation
  - We use WordNet as the source of word relations to calculate semantic similarity of words, but unfortunately, some words in the dataset are not included in WordNet.
  - We apply the “gestalt pattern matching” algorithm (Ratcliff et al. 1988) as a supplement, which computes the similarity of two strings as the number of matching characters divided by the total number of characters in the two strings.

- Parameters adjustment
  - the number of iterations in the weight calculation algorithm
  - the gap penalty in Needleman-Wunsch algorithm

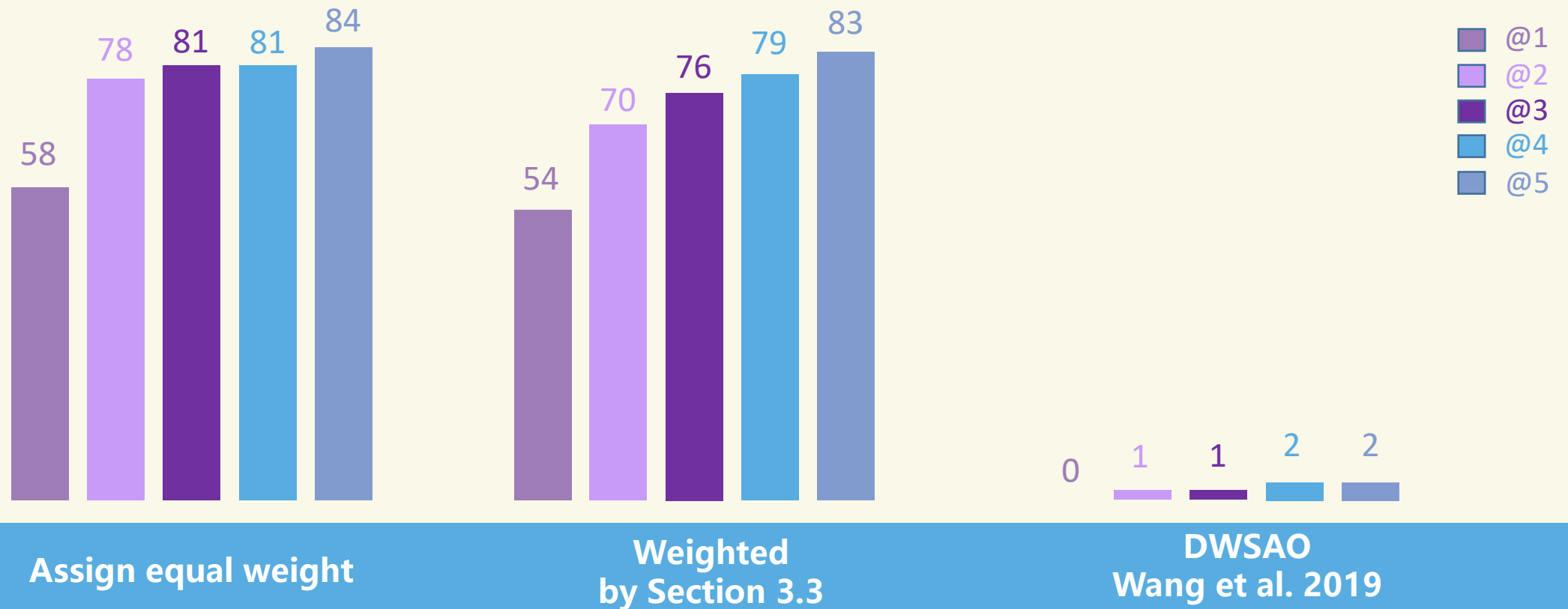
4.2 Experimental setup

- The number of iterations
  - One can determine whether it is stable by observing the trend of the weights after several iterations.
  - We randomly select 6 components to plot their trends with the number of iterations.
  - The number of iterations is fixed to 10 in this article



- The gap penalty
  - We choose multiple values for comparison, such as -0.05, -0.1, -0.15, -0.2 and -0.3.
  - The word alignment, patent similarity matrix and patent similarity will not be affected.
  - The gap penalty is set to -0.05 in this paper.

4.3 Experimental results and discussions



TOTAL 84 pairs

This figure shows the results of our methods and DWSAO method. Each patent is compared with other 1,009 patents, then patents of Top 1 (@1), Top 2 (@2), Top 3 (@3), Top 4 (@4) and Top 5 (@5) highest similarity are chosen to form 5 collections and then to judge how many of 84 pairs of patents are covered.

5 Conclusion

Conclusion

- This study proposes an improved semantic analysis for measuring patent similarity on the basis of entities and semantic relations (functional and non-functional relations), which takes semantic direction of each sequence structure and the word order information of each component into consideration.
- To verify the effectiveness and performance, a case study is conducted. The results show that our approach is significantly more accurate and is not sensitive to several core parameters

Thank You

THANKS FOR WATCHING