Wrocław, 13.06.2017

Wydział Informatyki i Zarządzania
Politechnika Wrocławska

Informacja o proponowanej do otwarcia rozprawie doktorskiej

Tytuł rozprawy: "Heterogeneous data in prediction of research topic dynamics" (pol. Dane heterogeniczne w predykcji dynamiki tematyk badawczych)

Doktorant: mgr inż. Rajmund Klemiński
Promotor: prof. dr hab. inż. Przemysław Kazienko
Promotor pomocniczy: dr. inż. Tomasz Kajdanowicz

Uzasadnienie podjęcia tematu

Analiza i próby predykcji rozwoju tematów badawczych są obszarem badań głęboko zakorzenionym w dziedzinie scjentometrii, jednak czerpiącym wielkie korzyści z progresywnego wzbogacania swoich metod zarówno o narzędzia statystyczne jak i uczenia maszynowego. Dziedzina frontów badawczych *(research fronts)* jest zdominowana przez metody sieciowe [1], w tym wykorzystanie cech topologicznych do przewidywania formujących się, szybko rosnących tematyk [2]. W tym ujęciu rozwój i przewidywanie odbywa się głównie za pomocą cytowań [3], jednakże proponowano metody oparte na innych miarach, jak np. przyrost informacji w tekstach [4], informacji bibliometrycznych [5] czy też analizy retorycznej [6]. Rozwiązania takie skupiają się na jednej grupie czy też rodzaju cech; jeżeli mapowanie nauki jest prowadzone za pomocą relacji między autorami [7] to nie dochodzi do porównania z innymi możliwymi mapami. Analogicznie ma się rzecz z cechami, co oznacza, że nisza badań nad hybrydowymi podejściami wykorzystującymi heterogeniczne dane nie jest zapełniona. Dodatkowo, proponowane rozumienie dynamiki jest w znaczący sposób łatwiejsze do interpretacji, pozwalając na użycie w naturalnie nasuwającym się celu: podejmowaniu decyzji o podjęciu bądź zarzuceniu badań nad daną tematyką w zależności od przewidywanego na przyszłe lata stanu.

Cel rozprawy

Celem rozprawy jest opracowanie metody fuzji heterogenicznych danych o publikacjach naukowych (tekst dokumentu, sieć społeczna autorów i cytowań, metadane) do zunifikowanej postaci. Z tak ujednoliconych danych pozyskiwane będą tematyki, tj. abstrakcyjne reprezentacje tematów naukowych poruszanych przez podzbiory publikacji. Ostatecznym celem rozprawy doktorskiej jest zaproponowanie metody uczenia rankingów *(learning to rank)* która, na bazie wcześniej wymienionych danych i obiektów, pozwala skonstruować wielowymiarowy ranking tematyk badawczych z następstwem czasowym, tj. przewidywać ranking tematyk dla przyszłych punktów w czasie.

Metodyka badań

W badaniach nacisk zostanie położony na eksperymenty na danych rzeczywistych, walidując wyniki przewidywania poprzez bezpośrednie porównanie z historycznymi danymi rzeczywistymi pozyskanymi z takich zbiorów danych jak DBLP [8] czy PubMed. Takie porównania posłużą zarówno do weryfikacji głównego celu rozprawy jak i porównania między sobą poszczególnych metod wchodzących w skład opracowywanego modelu.
Jakość uzyskiwanych reprezentacji tematyk zostanie zbadana metodami oceny jakości modelu takimi jak *„left-to-right"* [9] oraz miarami spójności tematyk [10].
Przeprowadzone zostaną również analityczne badania wpływu parametrów zastosowanych metod na wyniki predykcji, a także analiza statystyczna

Zakres rozprawy

W ramach pracy zaplanowano następujące zadania:

1. Pozyskanie dostępu do danych rzeczywistych
2. Stworzenie sieci autorów i cytowań
3. Opracowanie metody unifikacji danych heterogenicznych
4. Wykorzystanie modeli syntaktycznych i semantycznych do reprezentacji tematyk
5. Analiza i weryfikacja pozyskanych tematyk
6. Zaproponowanie metod uczenia do rankingowania kompatybilnych z danymi heterogenicznymi
7. Wykorzystanie zaproponowanego modelu do predykcji dynamiki rozwoju tematyk badawczych
8. Analiza wpływu parametrów zastosowanych metod na skuteczność predykcji
9. Eksperymenty weryfikacyjne na danych rzeczywistych

Uzyskane wyniki

Kluczowe cele rozprawy, tj. zaproponowanie metody unifikującej dane heterogeniczne oraz kompatybilnego z nią modelu nauczania rankingowania, wymagają dostępu do bogatych zbiorów danych, umożliwiających budowę sieci autorów i cytowań, a także wyodrębnienie meta-danych i samych dokumentów. W tym celu pozyskano dostęp do danych PubMed nad którymi badania rozpocząć się mają w niedalekiej przyszłości.

Przeprowadzone zostały badania na zbiorze DBLP z wykorzystaniem „bag of words" oraz LDA jako technik identyfikacji tematyk publikacji, bazując na tytułach tychże. Wyniki wykazały skuteczność przewidywania dynamiki tematyk badawczych w sensie ilości publikacji oraz wielkości społeczności badawczej, przy ograniczonej skuteczności dla miar przyrostu rok-do-roku. Dodatkowo zaproponowano modyfikację metody LDA dla zbioru tytułów publikacji naukowych i za jej pomocą wykazano, że specyfika języka wykorzystywanego w tytułach publikacji uniemożliwia wykorzystanie metod modelowania tematyk *(topic modeling)* na ich zbiorze. [11].

Literatura

1. Boyack, Kevin W., and Richard Klavans. "Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?." *Journal of the American Society for Information Science and Technology* 61, no. 12 (2010): 2389-2404.

2. Shibata, Naoki, Yuya Kajikawa, Yoshiyuki Takeda, and Katsumori Matsushima. "Detecting emerging research fronts based on topological measures in citation networks of scientific publications." *Technovation* 28, no. 11 (2008): 758-775.

3. Small, Henry. "Tracking and predicting growth areas in science." *Scientometrics* 68, no. 3 (2006): 595-610.

4. Lucio-Arias, Diana, and Loet Leydesdorff. "An indicator of research front activity: Measuring intellectual organization as uncertainty reduction in document sets." *Journal of the American Society for information Science and Technology* 60, no. 12 (2009): 2488-2498.

5. Mund, Carolin, and Peter Neuhäusler. "Towards an early-stage identification of emerging topics in science—The usability of bibliometric characteristics." *Journal of Informetrics* 9, no. 4 (2015): 1018-1033.

6. Prabhakaran, Vinodkumar, William L. Hamilton, Dan McFarland, and Dan Jurafsky. "Predicting the rise and fall of scientific topics from trends in their rhetorical framing." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2016.

7. Zhao, Dangzhi, and Andreas Strotmann. "Evolution of research activities and intellectual influences in information science 1996–2005: Introducing author bibliographic-coupling analysis." *Journal of the American Society for Information Science and Technology* 59, no. 13 (2008): 2070-2086.

8. Ley, Michael. "DBLP computer science bibliography." (2005).

9. Wallach, Hanna M., Iain Murray, Ruslan Salakhutdinov, and David Mimno. "Evaluation methods for topic models." In *Proceedings of the 26th annual international conference on machine learning*, pp. 1105-1112. ACM, 2009.

10. Newman, David, Jey Han Lau, Karl Grieser, and Timothy Baldwin. "Automatic evaluation of topic coherence." In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 100-108. Association for Computational Linguistics, 2010.

11. Klemiński, Rajmund and Kazienko Przemyslaw, „Identifying promising research topics in Computer Science", 4th European Network Intelligence Conference - 11.-12. September 2017 Duisburg, obecnie w recenzji

Podpis doktoranta

……