Abstract

This dissertation provides a comprehensive investigation of Reddit's information structure, through application of advanced Natural Language Processing (NLP) and graph network analysis. The research is reported in 7 contributions, each addressing various dimensions of Reddit's analysis.

Initially, the research has been focused on a thorough characterization of Reddit to delineate its topical landscape, coupled with extensive analysis across thousands of subreddits. In response to advancements in NLP, the study adopted various text processing methods to model Reddit's information structure based directly on the content of posts and text embeddings. Subsequent analysis utilized crossposting data to construct a directed graph network, which revealed critical insights into subreddit interactions and community structures. A novel method was introduced for detecting inter-subreddit similarities, using Named Entity Recognition (NER) and graph networks, offering a granular understanding of topical relationships and community dynamics. Recent advances in social network analysis, specifically the concepts of "bridges" and "gateways", were tested and expanded with the introduction of the new concept of "highways". Temporal network analysis, covering the period from 2015 to 2023, highlighted evolving thematic shifts and the consolidation of subreddit communities, while comparing the efficacy of the Louvain and Leiden methods in community detection. These novel approaches uncovered previously unknown relationships between subreddits.

Overall, the results contained in the Dissertation enhance the understanding of Reddit's ecosystem through innovative methodologies and detailed analyses at both micro and macro levels. They provide significant insights into online community dynamics and contribute to the advancement of NLP applications and graph analysis techniques, laying the groundwork for future research.

Keywords: Reddit, natural language processing, graph networks, social media

Streszczenie

Ta rozprawa przedstawia kompleksowe badanie struktury informacyjnej Reddita poprzez zastosowanie zaawansowanych metod przetwarzania języka naturalnego (ang. Natural Language Processing, NLP) oraz analizy sieci grafów. Badania opisano w 7 pracach, z których każda odnosi się do różnych aspektów analizy Reddita.

Początkowe badania koncentrowały się na dokładnej charakterystyce Reddita w celu zarysowania jego sfery tematyczne poprzez rozległą analizę tysięcy subredditów. W odpowiedzi na postępy w dziedzinie NLP, w badaniu zastosowano różne metody przetwarzania tekstu, aby modelować strukturę informacyjną Reddita bezpośrednio na podstawie treści postów i zanurzeń tekstu (ang. text embedding). Kolejne analize wykorzystały dane dotyczące krzyżowych publikacji, aby zbudować skierowaną sieć grafów, co ujawniło kluczowe spostrzeżenia dotyczące interakcji między subredditami i strukturami społeczności. Wprowadzono nowatorską metodę wykrywania podobieństw między subredditami, opartą na rozpoznawaniu nazwanych bytów (ang. Named Entity Recognition, NER) i sieciach grafów, co umożliwiło szczegółowe zrozumienie relacji tematycznych i dynamiki społeczności. Najnowsze postępy w analizie sieci społecznościowych, w szczególności koncepcje "mostów" (ang. bridges) i "bram" (ang. gateways), zostały przetestowane i rozszerzone poprzez wprowadzenie nowego pojęcia "autostrad" (ang. highways). Analiza sieci czasowej, obejmująca okres od 2015 do 2023 roku, podkreśliła ewoluujące zmiany tematyczne oraz konsolidację społeczności subredditów, porównując jednocześnie skuteczność metod Louvain i Leiden w wykrywaniu społeczności. Te nowatorskie podejścia odkryły wcześniej nieznane relacje między subredditami.

Wyniki zawarte w tej rozprawie poszerzają zrozumienie ekosystemu Reddita dzięki innowacyjnym metodologiom i szczegółowym analizom na poziomach mikro i makro. Dostarczają one istotnych obserwacji dynamice społeczności internetowych oraz przyczyniają się do rozwoju zastosowań NLP i technik analizy grafów, kładąc fundamenty pod przyszłe badania.

Słowa kluczowe: Reddit, przetwarzanie języka naturalnego, sieci grafowe, media społecznościowe