Kioto, Dnia 15go Września 2018

Recenzja Rozprawy Doktorskiej

Autor Rozprawy: Michał Jankowski-Lorek (Polsko-Japońska Akademia Technik Komputerowych)

Tytuł Rozprawy: "Models and Support Methods of the Wikipedia Knowledge Community"

Zawartość Rozprawy

Praca doktorską poświęcona jest w ogólnym zarysie badaniom nad Wikipedią. Szczegółowa problematyka badawcza pracy doktorskiej dotyczy metod oceniania jakości artykułów w Wikipedii oraz badaniu sieci społecznej autorów Wikipedii. Wikipedia jest obecnie największą encyclopedią która jest ogólnie dostępna i która zawiera kompedium wiedzy na większość możliwych tematów. Rozróżniającą characterystyką Wikipedii jest jej metoda tworzenia która jest oparta na pracy rzeszy woluntariuszy. Biorąc pod uwagę wielkość i użyteczność Wikipedii ważne jest jej dokładne zrozumienie i ciągłe ulepszenie. Praca jest poświęcona tej tematyce i składa się z pięciu głównych rozdziałów.

Rozdział pierwszy zawiera wstęp do rozprawy i przekazuje odpowiednią wiedzę czytelnikowi na temat charakteru całej pracy i jej zakresu. W podrozdziale 1.1 opisane są konkretne problemy naukowe jakimi zajmuje się autor w swojej pracy, natomiast podrozdział 1.2 zawiera opis literatury dotyczącej każdego z tych problemów naukowych. Dodatkowo autor wyjaśnia szczegółowo co brakuje w poprzedniej literaturze badawczej.

W drugim, trzecim i czwartym rozdziale autor streszcza tematy trzech problemów naukowych stanowiących podstawę pracy. Są one opisane poprzez zawarcie w pracy materiału opublikowanego wcześniej w trzech artykułach które ukazały się w międzynarodowych czasopismach naukowych.

W rozdziale drugim zatytuowanym "*Verifying Social Network Models of Wikipedia Knowledge Community*" autor analizuje i opisuje sieć społeczną Wikipedii prowadząc badania nad autorami Polskiej Wikipedii. Porównując deklarowane relacje między autorami Wikipedii z historią ich wzajemnej interakcji można zweryfikować kilka ważnych hipotez odnośnie działania całej Wikipedii i w szczególności charakteru jej sieci społecznej. Autor w tym rozdziale weryfikuje potencjalną interpretację relacji w sieci społecznej autorów wikipedii.

W rozdziale trzecim zatytuowanym "*What Makes a Good Team of Wikipedia Editors? A Preliminary Statistical Analysis*" autor prezentuje automatyczną metodę na analizę jakości

artykułów Wikipedii poprzez wykorzystanie informacji na temat wzajemnych relacji między autorami Wikipedii.

W rozdziale czwartym zatytuowanym *"Modeling Wikipedia Admin Elections using Multidimensional Behavioral Social Networks"* autor opisuje metodę na przewidywanie rezultatów elekcji administratorów Wikipedii. Także i tym razem wzajemne relacje między autorami Wikipedii są automatycznie rozpoznawane i użytkowane w celu przewidzenia czy dany autor będzie głosował na danego kandydata na administratora Wikipedii czy nie.

Rozdział piąty zawiera życiorys naukowy autora i jest końcowym rozdziałem pracy doktorskiej.

Zasadnicze Osiągniecia

Głównymi osiągnięciami pracy są trzy badane zagadnienia: a) modelowanie i zrozumienie społeczności Wikipedii, b) ocenianie jakości artykułów Wikipedii poprzez analizę autorów artykułów i c) modelowanie oraz predykcja rezultatów elekcji administratorów Wikipedii.

Pierwsze zagadnienie stanowi propozycja nowatorskiego sposobu analizy autorskiej społeczności Wikipedii. Autor stawia kilka pytań i tez na podstawie wcześniejszej powiązanej literatury i stara się je zweryfikować poprzez analizę behawioranej społecznej sieci oraz bezpośrednie zapytanie członków tej społeczności. Połączenie modelowania społeczności Wikipedii na podstawie aktualnych zachowań jej członków z informacjami deklarowanymi przez tych samych członków to ciekawy i oryginalny pomysł. Dzięki temu możliwe jest nie tylko lepsze zrozumienie w jaki sposób Wikipedia jest tworzona i zarządzana ale także odrzucenie kilku tez które były wcześniej często używane w powiązanej literaturze, szczególnie w celu oceny wiarygodności i jakości tworzonych artykułów. W tym sensie jest to najważniejsze osiągnięcie autora w jego pracy doktorskiej.

Drugie osiągnięcie to oryginalne rozwiązanie na oszacowanie jakości artykułów Wikipedii. Proponowana metoda na ocenę jakości artykułów Wikipedii ma dość wysoką efektywność I dodatkowo jest ona nowatorska. Choć niektóre wcześniejsze rozwiązania też były oparte na charakterystycznych właściwościach autorów Wikipedii, żadna że wcześniejszych prac nie proponowała użycia informacji na temat kompozycji, struktury i zasad współpracy między członkami zespołu który współtworzy ten sam artykuł Wikipedii.

Brak wiarygodnych ekspertów którzy mogliby być administratorami Wikipedii jest motywacja na kolejne badania których celem jest weryfikacja wpływu społecznych relacji na wyniki głosowania na administratorów Wikipedii. Autor analizuje czynniki jakie wpływają na fakt że dany członek społeczności Wikipedii będzie głosował za bądź przeciw danemu kandydatowi na administratora Wikipedii i poprzez analizę jest w stanie odpowiedzieć na ważne pytania dotyczące charakteru współpracy oraz relacji między autorami Wikipedii. Model na predykcję wyników głosowania jest też opisany. Biorąc pod uwagę stosunkowo małą ilość wcześniejszych prac poświęconych modelowaniu wyników głosowania na administratorów Wikipedii, jest to ważne osiągnięcie autora.

Praca stanowi spójną całość poprzez opisanie technologii które dotyczą Wikipedii i pozwalają nie tylko na a) zrozumienie i modelowanie wzajemnych relacji między członkami społeczności

Wikipedii ale też na b) określenie jakości artykułów i c) przewidywanie kto zostanie a kto nie administratorem Wikipedii.

Wkład do Dyscypliny Informatyki

Praca zawiera propozycję nowych automatycznych metod oceniania jakości tworzonego materiału w sieci, w szczególności w Wikipedii, oraz propozycję nowej metody analizy współpracy woluntariuszów w cely współtworzenia dobrej jakości treści, i jest poprzez to ściśle związana z dziedziną informatyki. Dodatkowo proponowane metody mogą też posłużyć przewidywaniu rezultatów przyszłych wyborów administratorów Wikipedii. Pozwalają one na lepsze zrozumienie i efektowne użytkowanie dużej ilości danych jakie można zebrać z zapisów interakcji autorów Wikipedii a poprzez to mogą pomoc w lepszej organizacji pracy w Wikipedii oraz jej zarządzaniu. Ponieważ Wikipedia jest najbardziej znaną stroną spolecznosciową sieci 2.0, dokonania autora w obecnej pracy mają szersze zastosowania dla lepszego zrozumienia I efektywnego zarządzania wielu przypadków społecznościowych sieci oraz stron sieci 2.0.

Oprócz osiągnięć w postaci nowych metod i algorytmów mocną stroną pracy są obszerne eksperymenty przeprowadzone na kilku różnych zbiorach danych pobranych z Wikipedii oraz od rzeczywistych użytkowników Wikipedii i odpowiedzi na szereg ważnych zagadnień dotyczących procesów jakie zachodzą w Wikipedii.

Omówienie Treści i Wyników Rozprawy

Uwagi Pozytywne

- Praca poświęcona jest ważnymi dziedzinami jakimi są analiza Wikipedii, zrozumienie procesów zachodzących pomiędzy woluntariuszami współtworzącymi treści w Wikipedii oraz automatyczna ocenę jakości treści Wikipedii i jakości administratorów.
- Proponowane metody analizy są prawidłowe i w dużej części są oparte na nowatorskich pomysłach. Pozwalają na lepsze zrozumienie Wikipedii i ogólnie także sieci społecznościowych. Proponowane metody są też przetestowane na rzeczywistych danych.
- Mocną stroną pracy jest to że zawiera ona oryginalne treści które powinny być użyteczne nie tylko w zakresie informatyki ale także w zakresie szerokiej gamy nauk społecznych i kognitywnych.
- Wszystkie proponowane metody są szczegółowo opisane i powiązane ze wcześniejszymi rozwiązaniami. Praca jest napisana w sposób jasny i przejrzysty a rezultaty są w wystarczający sposób dokładnie omówione.
- Wcześniejsza literatura która jest powiązana z problemami omawianymi przez autora została dobrze opisana i stanowi odpowiednie wprowadznie czytelnika do pokrewnych dziedzin. Dodatkowo autor wyjaśnia szczegółowo co brakuje w poprzedniej literaturze badawczej oraz podaję wytyczne odnośnie badań które mogą być prowadzone w przyszłości.
- Mimo że praca zawiera zawartości trzech osobnych publikacji jako większość materiału, autor zamieścił obszerne wprowadzenie i wytłumaczenie treści w pierwszym rozdziale. Ten sposób powoduje że rozdziały są bardziej spójne i łatwiejsze w zrozumieniu.
- Autor rozpoznaje i wyjaśnia obszernie problemy oraz niedoskonałości swoich metod badawczych np., niemożliwości zapytania całkowicie losowej grupy autorów Wikipedii. To pozwala na lepsze zrozumienie wyników i osiągnięć autora oraz daje możliwość poprawy obecnych badań bądź ich dalszego rozszerzenia w przyszłości.

Uwagi Krytyczne i Dyskusyjne

1. Choć trzy naukowe problemy którymi autor się zajmuje różnią się znacząco, rozwiązania proponowane przez autora są dość zbliżone do siebie. Wszystkie metody są oparte na użyciu standardowych klasyfikatorów i mają podobne atrybuty obliczone z wejściowych danych w oparciu których podejmowane są automatycznie decyzje klasyfikowania.
2. Autor często wspomina że jego metody bazują na grafach, jednak praktycznie metody z teorii grafów są używane w raczej małym stopniu. Poza analizowaniem triadów, proponowane rozwiązania nie mają raczej za wiele wspólnego z grafami chociaż charakter danych pozwalałaby na zastosowanie bardziej skomplikowanych metod z teorii grafów w celu ich lepszego zrozumienia bądź przewidywania.
3. W niektórych rozdziałach pracy, zwłaszcza w rozdziale drugim, przydałoby się użyć bardziej formalnych opisów w sensie matematycznym raczej niż używając samego tekstu. Także lepiej jest uporządkować rezultaty w tabelkach niż tylko wspominać je w głównym tekście.
4. Niektóre hipotezy analizowane przez autora dotyczą raczej zagadnień kognitywnych lub nawet psychologicznych jako że bezpośrednio wiążą się one z długością bądź obszernością pamięci autorów Wikipedii. By je zbadać trzeba analizować po jak długim czasie, dana osoba będzie pamiętała jakich innych użytkowników, a to już nie jest zagadnienie ściśle powiązane tylko z Wikipedia i z jej charakterem lecz także z szerszymi zagadnieniami kognitywnymi. Wobec tego, albo niektóre z hipotez powinny być zmodyfikowane tak by efekt upływu czasu oraz efekty zapamiętywania były wzięte pod uwagę i poprzez to analiza byłaby bardziej specyficzna dla samej Wikipedii, albo literatura z zakresu właściwości ludzkiej pamięci i procesów zapamiętywania/zapominania powinna być wspomniana i omówiona, na przykład, „recency/primacy effects" lub odkrycia dokonane przez Ebbinghausa (np. „forgetting curve") mogłyby być wspomniane w pracy.
5. Wobec dość dużego nacisku na badania procesu zapamiętywania innych autorów jaki ta praca doktorska kładzie, dość zaskakujący jest fakt że autor nie rozważa otwarcie kwestii „czasu" w swoich badaniach jako ważnego elementu oraz atrybutu. Autorzy i administratorzy Wikipedii którzy brali udział w badaniach naturalnie różnią się pod względem okresu jaki spędzili edytując Wikipedię oraz częstotliwości z jaką pracowali bądź nadal pracują nad artykułami. Praca w obecnym kształcie nie zawiera bezpośrednich odniesień do czasu, i raczej nie jest on bezpośrednio brany pod uwagę jako jeden z atrybutów dla klasyfikatorów poza kilkoma wyjątkami. Dodatkowo czas powinien być wzięty pod uwagę w procesie przewidywania jakości artykułów jaki jest opisany w rozdziale trzecim, gdyż, intuicyjnie, artykuł stworzony niedawno ma mniejsze szansę by być dobrej jakości niż starszy artykuł który ma większą szansę być wyedytowany przez dużą ilość autorów.
6. W rozdziale drugim (sekcja 5.5) autor analizuje dwa pytania/hipotezy (Q7 i Q8). Nie do końca jest oczywiste dlaczego potrzebna była binaryzacja wyników i dlaczego metody regresji (np. linear regression) nie zostały użyte zamiast klasyfikacji.
7. Dane użyte w rozdziale czwartym, choć autentyczne, są jednak dość stare (zebrane koło 10 lat temu). W obecnym czasie charakter pracy w Wikipedii mógł się zmienić w porównaniu z latami 2005-2010. Autor powinien omówić w pracy kwestie tego czy dane są nadal reprezentatywne.
8. Edycje autorów Wikipedii są traktowane w ten sam sposób. Jednak intuicyjnie wiadomo że niektóre edycje są ważniejsze a inne mniej ważne. Na przykład zmiana (np. dodanie jednego zdania) w pierwszym akapicie jest typowo ważniejsza niż taka sama zmiana dokonana na samym końcu artykułu. Także zmiana dotycząca dużej ilości tekstu jest intuicyjnie ważniejsza

od zmiany dotyczącej małej ilości tekstu. Autor powinien wytłumaczyć dlaczego pozycja edycji oraz ich wielkość nie zostały wzięte pod uwagę w jego badaniach.

9. Dodatkowo autor wziął pod uwagę pojedyncze słowa lub zbiory słów dla rozróżnienia różnych rodzajów edycji (moving/additions/deletions). Być może operowanie na poziomie zdań byłoby lepszym rozwiązaniem tutaj co tez powinno być wytłumaczone.
10. Część parametrów używanych w pracy np. *max_recent*, *distance_cutoff*, *discussion_distance* i inne została wyznaczona raczej arbitralnie. Lapiej byłoby umotywować dokładnie wybór wartości paramaterów, na przykład, na podstawie analizy danych bądź na podstawie wcześniejszych rozwiązań.

Uwagi Redakcyjne i Językowe

- "multidimensional implicit social netowrks (MBSN)" w rozdziale 1.2.1 powinno być zamienione na "multidimensional behavioral social networks (MBSN)"
- Dobrze byłoby usunać oryginalne numery stron z załączonych artykułów jako że numeracja stron powinna być taka sama dla całej pracy doktorskiej
- Sekcja 4.2.3 zawiera to samo zdanie powtórzone dwa razy: "An average Wikipedian has 14.5 neighbours..."
- "Hypotheses 25" powinno być zamienione na "Hypotheses 2 and 5" w rozdziale pierwszym
- "two factions" powinno być zmienione na "two fractions" na stronie 7 w rozdziale trzecim
- Atrybuty takie jak *team_id_1* lub *disc_nidc* i inne w rozdziale trzecim nie są zdefiniowane, i czytelnik musi domyślać się co one faktycznie reprezentują

Podsumowanie

Reasumując stwierdzam że Pan Michał Jankowski-Lorek posiada ogólną wiedzę teoretyczną i praktyczną w zakresie analizy systemów spolecznosciowych opartych na zasadzie wolnej treści, a w szczególności Wikipedii, i w zakresie automatycznego oceniania jakości materiałów zawartych w Wikipedii oraz analizy charakteru współpracy między autorami Wikipedii.

Praca poświęcona jest trzem precyzyjnie sformułowanym problemom i zawiera wiele znaczących elementów oryginalnych. Nie znalazłem w pracy większych błędów merytorycznych a jakiekolwiek zastrzeżenia mają generalnie charakter dyskusyjny.

Proponowane rozwiązania zostały zweryfikowane przez akademickie środowisko poprzez publikacje w międzynarodowych czasopismach i prezentację na konferencjach oraz warsztatach w dziedzinie analizy informacji i badań nad sieciami społecznymi. Badania zostały przeprowadzone z użyciem rzeczywistych danych.

Praca spełnia wymagania zdefiniowane przez Artykuł 13 ustawy z dnia 14 marca 2003-go o stopniach naukowych i tytule naukowym (z późniejszymi poprawkami). Jako że praca stanowi oryginalne rozwiązanie konkretnego i ważnego problemu naukowego, które zostało też rzetelnie zweryfikowane, rekomenduję przyjęcie rozprawy i dopuszczenie Pana Michała Jankowski-Lorka do publicznej obrony.

Adam Jatowt