Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji

Kierunek: **Informatyka algorytmiczna (INA)**
Specjalność:

PRACA DYPLOMOWA INŻYNIERSKA

Zastosowanie szkiców danych w analizie dużych grafów

Application of data sketches in the analysis of large graphs

Paweł Polerowicz

Opiekun pracy
dr inż. Jakub Lemiesz

Słowa kluczowe: TODO, TODO, TODO

WROCŁAW, 2024

Streszczenie

Polski

Słowa kluczowe: TODO, TODO, TODO

Abstract

English

Keywords: TODO, TODO, TODO

Spis treści

Spis rysunków

Spis tabel

Spis listingów

Skróty

ISO (ang. *International Standards Organization*)

Rozdział 1

Wstęp

TODO: Tutaj będzie bardzo ogólne wprowadzenie do pracy.

- Motywacja
- Krótki i "wysokopoziomowy"opis problemu
- Podsumowanie osiągnięć pracy

1.1. Struktura pracy

TODO: Kilka(dziesiąt) słów o strukturze i zawartości pracy. Omówienie po kolei rozdziałów i ewentualnych dodatków. Bardzo skrótowo, bo wszystko będziemy potem i tak rozwijać. Coś w jak niżej (do dopracowania).

Pierwszy rozdział stanowi niniejszy wstęp, przedstawiający ogólny zarys problematyki pracy i skrótowo podsumowujący jej wkład badawczy. W drugim rozdziale znajduje się opis problemu wraz z formalną definicją i przedstawieniem różnych jego wariantów. Przedmiotem trzeciego rozdziału jest przegląd literatury związanej z analizą wielkich grafów, z podziałem na zastosowane metodyki oraz tabelą ilustrującą porównanie znanych struktur i algorytmów. W czwartym rozdziale szczegółowo omówione zostały szkice danych, od ich formalnej definicji do bardziej praktycznych przykładów ich wykorzystanie, także w kontekście niniejszej pracy. Piąty rozdział zawiera właściwy opis tego, co zostało zrobione [TODO: przepisać nieco bardziej szczegółowo]. W rozdziale szóstym opisane zostały przeprowadzone eksperymenty, wraz z prezentacją wyników oraz wnioskami z nich płynącymi. Ostatni, siódmy rozdział, stanowi podsumowanie pracy. Zawarte zostały w nim ogólne konkluzje na temat pracy oraz możliwe kierunki dalszych badań. Pracy towarzyszy wykaz literatury oraz dodatek, zawierający opis dołączonej płyty CD [TODO: czy nadal załączamy płytę?] i instrukcję użytkowania części implementacyjnej.

Rozdział 2

Opis problemu

2.1. Analiza grafów

2.2. Główne sposoby modelowania problemu

2.3. Definicja formalna

Rozdział 3

Przegląd literatury

3.1. Rozmiar i układ treści na stronach dokumentu

3.2. MDL

Sketch synopses - CountMin, gSketch Graph summaries - Graph pattern matching over streams Graph stream algorithms Distributed graph systems

3.3. Metody oparte na modyfikacji macierzy sąsiedztwa

Jednym z najbardziej popularnych i być może najprostszym koncepcyjnie sposobem reprezentacji grafu jest macierz sąsiedztwa. Jej wiersze i kolumny odpowiadają poszczególnym wierzchołkom, a w komórkach przechowywane są wagi krawędzi pomiędzy nimi, o ile takowe krawędzie istnieją. W przypadku grafów nieważonych może to być np. wartość logiczna indykująca istnienie krawędzi lub ustalona stała. Ten sposób reprezentacji ma niewątpliwe zalety takie jak prostota implementacji i, przede wszystkim, stały czas dostępu do wag krawędzi. Z tego powodu

Jedną z pierwszych realizacji tej idei o przechowywania informacji o strumieniowanym grafie jest struktura TCM[5]. Ma ona postać macierzy o boku długości m, gdzie m jest pewną stałą. Podobnie jak w klasycznej macierzy sąsiedztwa, w jej komórkach składowane są wagi krawędzi. Zasadniczą różnicą jest natomiast sposób wyznaczania rzędu i kolumny odpowiadających danej parze wierzchołków. Są one bowiem wyznaczane przez wynik funkcji haszującej $H : V \rightarrow [1..m]$. Czas obliczania hasza jest stały, a co za tym idzie, złożoność czasowa zapytań i dodawania nowych krawędzi również. Teoretyczna złożoność pamięciowa także jest stała i wynosi $O(m^2)$. W praktycznych zastosowaniach wybór m zależy jednak często od liczby krawędzi i przejmuje się najczęściej m rzędu $O(\sqrt{|V|})$. Dokładność rezultatów zależy od rozmiaru macierzy i może być niska ze względu na kolizje haszy. Łatwo zauważyć, że jeśli m jest istotnie mniejsze od $|V|$ to może do nich dochodzić często, co powoduje traktowanie różnych krawędzi jako kolejnych instancji tego samego połączenia. Autorzy, świadomi tego ograniczenia, proponują zastosowanie kilku parami niezależnych funkcji haszujących i stworzenie na ich podstawie wielu szkiców grafu. Przykładowo, jeśli badaną zmienną jest suma wag kolejnych instancji krawędzi między danymi dwoma wierzchołkami, to algorytm może sprawdzić odpowiednie komórki dla wszystkich szkiców, a następnie zwrócić maksymalną wartość. Podejście to pozwala na analizę większych grafów niż w przypadku pojedynczego szkicu, ale ostatecznie nie rozwiązuje całkowicie problemu. Użyteczność struktury TCM w bazowej formie jest dyskusyjna, stanowi ona jednak punkt wyjściowy dla bardziej zaawansowanych rozwiązań.

Strukturą opartą na koncepcie podobnym do TCM jest *Graph Stream Sketch* (GSS)[3]. Celem autorów było stworzenie metody oferującej lepszą skalowalność dla wielkich grafów. Podobnie jak w TCM, funkcja haszująca mapuje zbiór wierzchołków na pewien mniejszy zbiór M-elementowy. Rozmiar macierzy jest natomiast równy m, $m < M$. Główną zmianą jest wprowadzenie dodatkowych cech opisujących wierzchołki. Na podstawie hasza $H(v)$ wyznaczany jest podpis wierzchołka $f(v)(0 \leq f(v) < F)$, gdzie $M = m \times F$ i $f(v) = H(v)\%F$, a także adres $h(v) = \lfloor \frac{H(v)}{F} \rfloor$. Adresy służą do wyznaczania rzędu i kolumny komórek. Komórki te mają one postać krotki lub, bardziej obrazowo, kubełka, w którym przechowywana jest para podpisów wierzchołków tworzących krawędź oraz kumulatywna waga krawędzi. Przechowywanie podpisów w komórkach pozwala zredukować ryzyko konfliktu haszy. Łatwo bowiem zauważyć, że nawet jeśli dwa różne wierzchołki mają taki sam adres, to istnieje duża szansa, że ich podpisy są różne. Z tego względu nowa krawędź jest dodawana do kubełka tylko w wypadku, gdy jest on pusty lub gdy istniejące w nim podpisy są zgodne z podpisami wierzchołków krawędź tą tworzących. W przeciwnym przypadku jest ona zapisywana w dodatkowym buforze, mającym postać listy sąsiedztwa pełnych haszy. Pozwala on na dodawanie nowych krawędzi z niskim ryzykiem kolizji, nawet jeśli sama macierz jest już zapełniona. Należy natomiast zauważyć, że część macierzowa struktury jest bardziej efektywna czasowo, oferując stały czas odpowiedzi na zapytanie, podczas gdy dla bufora jest on liniowy względem liczby wierzchołków. Dokładność odpowiedzi w części macierzowej zależy od długości podpisów. Potencjalnym problemem GSS jest niskie wykorzystanie pamięci w macierzy. Przy kolizji adresów nowe krawędzie mogą trafiać do bufora, mimo, że w samej macierzy pozostaje wiele pustych komórek. Aby temu zaradzić, autorzy proponują haszowanie krzyżowe (*ang. square-hashing*). Zakłada ona obliczanie dla każdego wierzchołka sekwencji niezależnych adresów. Podczas wstawiania nowych krawędzi algorytm sprawdza nie jedną komórkę macierzy, a kilka, zgodnie z sekwencją adresów i wybiera pierwszą spełniającą wymagania co do zgodności podpisów. Istnieje probabilistyczne ograniczenie na błąd względny zapytań postaci $Pr(\tilde{f}(s,d) - f(s,d)/\bar{w} > \delta) \leq \frac{|E|}{\delta m^2 4^f}$, gdzie $\tilde{f}(s,d)$ jest zwróconą sumą wag krawędzi (s,d), $f(s,d)$ jej rzeczywistą wartością, $\bar{w}$ średnią wagą krawędzi, a f długością podpisu.

Większość struktur służących podsumowujących strumieniowane grafy nie przechowuje informacji o czasie wystąpienia krawędzi. Nie wspierają one więc zapytań z zakresem czasowym, a więc np., czy dana krawędź wystąpiła w zakresie $[t, t + L)$. Tego typu zapytania mogą być kluczowe np. w przypadku analizy danych dotyczących rozprzestrzeniania się wirusów (TODO: Citation needed). Problem ten podejmuje praca proponująca strukturę Horae[2]. W jej wypadku krawędź $e_i = (s_i, d_i, w_i, t_i)$ jest wstawiana do komórki o adresie $(h(s_i|\gamma(t_i)), h(d_i|\gamma(t_i)))$, gdzie $\gamma(t_i) = \lfloor \frac{t_i}{gl} \rfloor$ i gl jest długością przedziałów czasowych. Intuicyjnie, zapytanie o pojawienie się krawędzi w zakresie czasowym $[T_b, T_e]$ moze być transformowane w sekwencję zapytań o pojedyncze zakresy, których wyniki są sumowane, a więc $Q([T_b, T_e]) = Q([T_b]), Q([T_{b+1}]), \ldots Q([T_e])$. Jednak dla takiego algorytmu złożoność czasowa jest liniowa względem liczby zakresów. Autorzy starają sie poprawić ten aspekt, zauważajac, iż przedział długości L może zostać zdekomponowany do co najwyżej $2 \log L$ podprzedziałów podsiadających dwie szczególne cechy. Po pierwsze, wszystkie zakresy czasowe w danym podprzedziale mają wspólny prefiks binarny. Po drugie, prefiksy różnych podprzedziałów mają różne długości. Z tego względu Horae zapamiętuje $O(\log(T))$ identycznych skompresowanych macierzy, gdzie T jest liczbą rozróżnialnych zakresów czasowych. Każda z nich jest utożsamiana z jedną warstwą struktury. Warstwy odpowiadają z kolei różnym długościom prefiksów. Dzięki temu zamiast wykonywać liniową względem długości przedziału czasowego liczbę zapytań, wystarczy zdekomponować przedział na podprzedziały i na ich podstawie wykonać co najwyżej jedno zapytanie na warstwę.

Metody oparte na macierzach w większości przypadków nie czynią założeń co do struktury grafu. Takie ogólne podejście oczywiście zapewnia wysoką uniwersalność, jednak w niektórych przypadkach może być nieefektywne. Przykładowo, jeśli wierzchołki w grafie są mocno zróżnicowane pod względem stopnia, a więc bardziej obrazowo, da się wyróżnić obszary gęste i rzadkie w grafie, to kolizje haszy mogą zdarzać się często. Struktura Scube[1] używa probabilistycznego zliczania do identyfikacji wierzchołków wysokiego stopnia. Przeznaczane jest dla nich więcej kubełków w macierzy niż dla wierzchołków o niskich stopniach, co pozwala bardziej efektywnie zarządzać zapełnieniem macierzy.

Metody takie jak GSS czy Horae, choć często dają przyzwoite wyniki przy odpowiednim dobraniu parametrów do badanego grafu, to ostatecznie cierpią z uwagi na ograniczoną skalowalność. Jedną z prób odpowiedzi na ten problem jest struktura AUXO[4]. Korzysta ona z macierzy przechowujących podpisy wierzchołków, podobnie jak GSS. Jednak, zamiast wstawiać nadmiarowe krawędzie do bufora o liniowym czasie dostępu, AUXO wykorzystuje wiele macierzy ustawionych w strukturę drzewa. Konkretnie, jest to binarne lub czwórkowe drzewo prefiksowe, w którego strukturę zaszyte zostały prefiksy podpisów wierzchołków. W ten sposób na każdym kolejnym poziomie drzewa podpisy przechowywane w komórkach mogą być coraz krótsze, gdyż informacja ta jest wbudowana w kształt struktury. Pozwala to osiągnąć logarytmiczny względem liczby krawędzi czas odpowiedzi na zapytania. Warto zauważyć, że złożoność pamięciowa jest ograniczona przez długość podpisów, która wyznacza maksymalną głębokość drzewa. Niemniej jednak liczba możliwych do przetworzenia krawędzi jest eksponencjalna w stosunku do liczby bitów podpisu, więc stosunkowo łatwo można dobrać wystarczające wartości. W praktycznych zastosowaniach autorzy wskazują, nieco niefortunnie, na złożoność pamięciową zbliżoną asymptotycznie do $O(|E|(1 - \log(E)))$. Jak widać, AUXO osiąga ekektywność pamięciową i skalowalność kosztem zwiększenia złożoności czasowej, co może być potencjalną wadą tego rozwiązania.

3.4. Embeddings

Rozdział 4

Szkice danych

4.1. Definicja

Rozdział 5

Główny przedmiot pracy

5.1. Co nas najbardziej interesuje

5.2. Jaki jest ogólny pomysł

5.3. Jak to dokładnie zrobiliśmy

Rozdział 6

Analiza wyników

6.1. Architektura eksperymentów

6.2. Wyniki

6.3. Wnioski

Literatura

[1] M. Chen, R. Zhou, H. Chen, H. Jin. Scube: Efficient summarization for skewed graph streams. *2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS)*, 2022.

[2] M. Chen, R. Zhou, H. Chen, J. Xiao, H. Jin, B. Li. Horae: A graph stream summarization structure for efficient temporal range query. *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, 2022.

[3] X. Gou, L. Zou, C. Zhao, T. Yang. Fast and accurate graph stream summarization. *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, 2019.

[4] Z. Jiang, H. Chen, H. Jin. Auxo: A scalable and efficient graph stream summarization structure. *Proceedings of the VLDB Endowment*, 16(6):1386–1398, 2023.

[5] N. Tang, Q. Chen, P. Mitra. Graph stream summarization. *Proceedings of the 2016 International Conference on Management of Data*, 2016.