Eksploracja zasobów internetowych

Wykład 1

Badanie struktury sieci WWW

mgr inż. Maciej Kopczyński

Białystok 2014

Rys historyczny

Idea sieci Web stworzona została w 1989 przez Tima Bernersa-Lee z CERN jako struktura do przechowywania, publikacji i wymiany informacji.

Jego głównym celem było skoordynowanie prac na projektami naukowymi poprzez ułatwienie wymiany wyników badań pomiędzy zespołami badawczymi.

Rys historyczny

Tim Berners-Lee stworzył pierwszy pakiet narzędzi do działania sieci WWW – przeglądarkę, serwer oraz strony internetowe.

Na rozwój sieci WWW duży wpływ miał jej otwarty i bezpłatny charakter w odróżnieniu od innych systemów wymiany informacji.

Należy pamiętać, że sieć WWW to nie jest Internet!

Rys historyczny

Pierwszą graficzną przeglądarką internetową był Mosaic, który powstał w 1993 roku.

Przełomem w rozwoju sieci WWW było połączenie jej z siecią Internet.

Sieć WWW jest archiwizowana między innymi przez instytucję Internet Archive, która działa od 1996 roku.

Cechy sieci WWW

Najbardziej istotne cechy sieci WWW to:

- struktura grafowa,
- węzły reprezentujące dokumenty, informacje, osoby, itd.,
- krawędzie jako powiązania między nimi,
- krawędzie są jednokierunkowe.

Rozwój sieci

Uniwersalność rozwiązania wpłynęła na szybki rozwój sieci:

- po 10 latach:
  - 150 milionów stron,
  - 1.7 miliardów linków,
- po 20 latach:
  - 4 miliardy stron,
  - 1 milion nowych stron dodawanych codziennie.

Szybkość rozwoju sieci wzrasta w czasie.

Rozwój sieci

Sieć WWW jest obecnie jedną z największych baz danych zawierającą wiedzę z praktycznie wszystkich dziedzin życia oraz nauki.

Bardzo istotna jest możliwość wyszukiwania wiedzy wśród tych danych. Bez dobrych mechanizmów wyszukiwania nie jest możliwe korzystanie z wartościowych informacji znajdujących się w sieci.

Przyszłość sieci jako bazy wiedzy

Przyszłość sieci jako źródła wiedzy uzależniona jest od możliwości oraz prostoty jej pozyskiwania.

Trzy główne możliwości:

- silniki przeszukiwania sieci (sieć istniejąca),
- *Topic Directories* (sieć istniejąca),
- *Semantic Web* (sieć przebudowana).

Topic directories

Topic directories są strukturami zawierającymi strony WWW pogrupowane zgodnie z ich zawartością. Cechy rozwiązania:

- struktura hierarchiczna,
- nie wymaga przebudowy dokumentów sieci,
- tworzone ręcznie,
- największe repozytorium: *dmoz.org*

Semantic web

Semantic web jest techniką dodawania metadanych opisujących zawartość dokumentu do poszczególnych stron WWW. Rozwiązanie zostało zaproponowane przez konsorcjum *w3c*.

Cechy rozwiązania:

- metadane są niewidoczne dla człowieka,
- opisy ułatwiają klasyfikację treści dokumentów,
- wymaga przebudowy dokumentów WWW,
- w połączeniu z rozwiązaniami SI umożliwia udzielanie odpowiedzi na zadane pytania.

Silniki wyszukiwania

Silniki wyszukiwania danych pozwalają na zwracanie kolekcji dokumentów zgodnych z zapytaniem złożonym ze słów kluczowych. Cechy rozwiązania:

- nie wymaga przebudowy dokumentów sieci WWW,
- pozwala na zwracanie wyników z uwzględnieniem istotności poszczególnych dokumentów,
- istotność obliczana na bazie słów kluczowych oraz połączeń pomiędzy stronami.

Silniki wyszukiwania

Pierwszą wyszukiwarką internetową był *Archie* stworzony w roku 1990. Obecnie, najwięksi dostawcy silników wyszukiwania danych w sieci WWW to:

- Google,
- Yahoo!,
- Bing,
- Altavista.

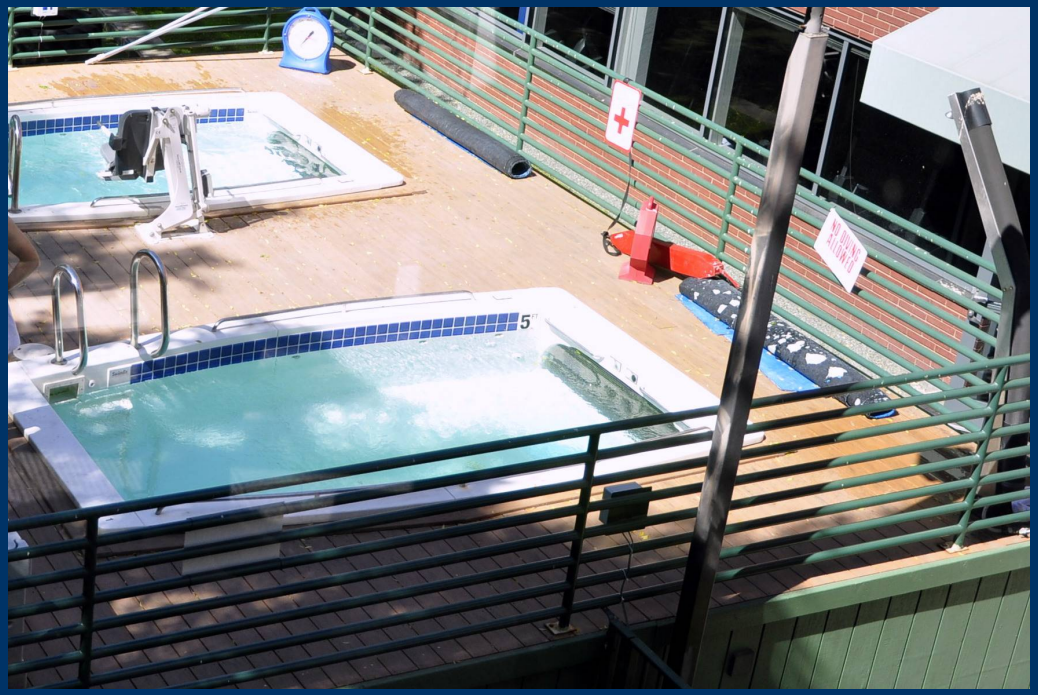
Kilka słów na temat Google

Google powstało w 1998 r. z pomysłu dwóch doktorantów Stanford University: Larry'ego Page'a oraz Sergey'a Brina.

Obecne fakty:

- ponad 54 000 pracowników,
- ok. 10 mld $ przychodu,
- główna siedziba: Mountain View, CA, USA,
- w Polsce: Kraków, Wrocław, Warszawa.

Kilka słów na temat Google




splunk

Silniki wyszukiwania

Głównymi elementami składowymi wyszukiwarki internetowej są:

- crawler przeszukujący sieć WWW,
- indekser przetwarzający pobrane dane,
- baza danych do zapisania pobranych danych,
- moduł wyszukiwania i sortowania dokumentów zgodnych z zapytaniem użytkownika,
- interfejs prezentujący zwrócone wyniki.

Crawlery

Celem crawlerów jest analiza struktury sieci WWW i budowanie na jej podstawie grafu oraz zachowywanie zawartości poszczególnych dokumentów.

Podstawowymi elementami składowymi crawlerów są:

- parser języka HTML,
- algorytmy analizy danych,
- indekser odwiedzonych stron.

Zapytania

Dużym problemem w definiowaniu zapytań przez użytkownika jest rozbieżność pomiędzy tym, co użytkownik chce odnaleźć a tym, w jaki sposób definiowane jest zapytanie.

Jedną z metod radzenia sobie z problemem jest generowanie słów-podpowiedzi.

Statystyki pokazują, że średnia długość zapytania to 2 – 3 słowa.

Crawlery – trudności

Problemy podczas analizy sieci WWW:

- opóźnienia sieciowe,
- tłumaczenie adresów domenowych na adresy IP dla każdej strony,
- poruszanie się po linkach tworzących pętle,
- rozmiar gromadzonych danych,
- stale zmieniająca się sieć,
- błędy działania serwerów, kod niezgodny z HTML.

Opóźnienia sieciowe

Nawiązanie połączenia TCP/IP z serwerem WWW wprowadza pewne opóźnienie.

Metodami radzenia sobie z tym problemem są:

- wielowątkowość,
- wykorzystanie wielu komputerów,
- programowanie asynchroniczne (zdarzenia, gniazda nieblokujące),

Adresy DNS

Serwer WWW jest dostępny pod adresem domenowym wymagającym translacji do adresu IP. Wielokrotna translacja tego samego adresu wprowadza opóźnienia.

Metodą radzenia sobie z tym problemem jest wykorzystanie mechanizmu cache'u.

Pętle wynikające z linków

Przeszukując sieć WWW często można trafić na pętle, co powoduje powrót do dokumentu początkowego. Tego typu sytuacja może „zawiesić” działania crawlera.

Metodą radzenia sobie z tym problemem jest wykorzystanie bazy odwiedzonych dokumentów (np. technika haszowania).

Rozmiar gromadzonych danych

Sieć WWW jest największą bazą wiedzy dostępną dla każdego. W 2005 roku Yahoo! poinformowało, że jego baza zaindeksowanych stron liczy 20 000 000 000 stron. Dostęp do zgromadzonych danych musi być realizowany natychmiastowo.

Metodami radzenia sobie z tymi problemami są:

- maksymalne przetworzenie pobranych danych,
- kompresja danych,
- wykorzystanie farm serwerów.

Stały rozwój sieci

Sieć WWW jest strukturą ewoluującą z ciągle dodawanymi nowymi stronami oraz zmieniającymi się istniejącymi dokumentami. Dobra wyszukiwarka internetowa musi zwracać aktualne wyniki.

Metodą radzenia sobie z tym problemem ciągłe przeszukiwanie sieci.

Niezgodność z kodem HTML

Nie każdy autor strony WWW umieszczonej w sieci jest programistą, stąd wiele dokumentów HTML jest niezgodnych z ustalonymi standardami. Parser kodu HTML musi być odporny na błędy w kodzie HTML poszczególnych stron.

Metodą radzenia sobie z tym problemem jest stworzenie niezawodnego parsera kodu HTML odpornego na błędy w składni dokumentu HTML.

Błędy działania serwerów WWW

Serwery WWW pracują z wykorzystaniem różnego oprogramowania, które może zawierać błędy powodujące ich niestabilne działanie. Crawler musi stabilnie działać nawet w przypadku niepoprawnej pracy serwera WWW.

Metodą radzenia sobie z tym problemem jest implementacja crawlera jako programu odpornego na wszelkie możliwe przypadki działania.

Sieć WWW jako graf

Struktura sieci WWW może być reprezentowana jako graf skierowany. Problemem jest jednak stworzenie tego grafu.

Sposoby reprezentacji grafu:

- macierz sąsiedztwa,
- listy incydencji.

Metody przeszukiwania sieci WWW i budowania grafu:

- wszerz,
- w głąb.

Macierz sąsiedztwa

Każda komórka macierzy jest opisana zależnością:

$$G[i,j]=\begin{cases}1 & \text{jeśli występuje krawędź } i,j\\ 0 & \text{jeśli nie występuje krawędź } i,j\end{cases}$$

Przykładowa macierz sąsiedztwa *G* opisująca graf:

	1	2	3	4	5
1	0	1	1	0	0
2	0	0	1	0	1
3	0	0	0	1	0
4	0	0	0	0	0
5	0	0	0	0	0

Złożoność pamięciowa: n^2

Listy incydencji

Listy incydencji mogą być zaimplementowane w oparciu o dynamiczne struktury danych.

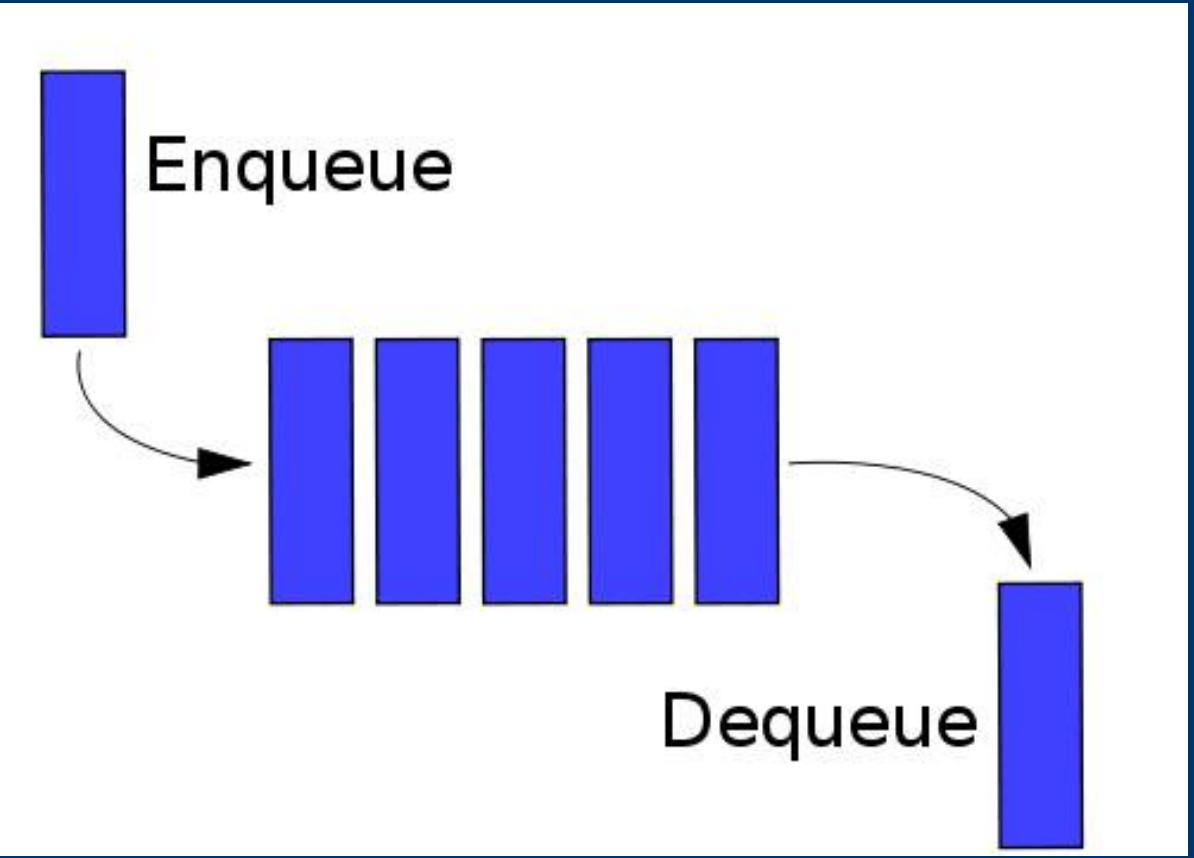
Przykładowa lista incydencji opisująca graf:

- T[1]: 2, 3
- T[2]: 3, 5
- T[3]: 4
- T[4]: ∅
- T[5]: ∅

Złożoność pamięciowa: $n + m$

Przeszukiwanie wszerz

Metoda przeszukiwania wszerz oparta jest o strukturę FIFO.



Przeszukiwanie wszerz

Cechą przeszukiwania wszerz jest równomierne odwiedzanie wszystkich odkrytych ścieżek.

Wykorzystywane oznaczenia:

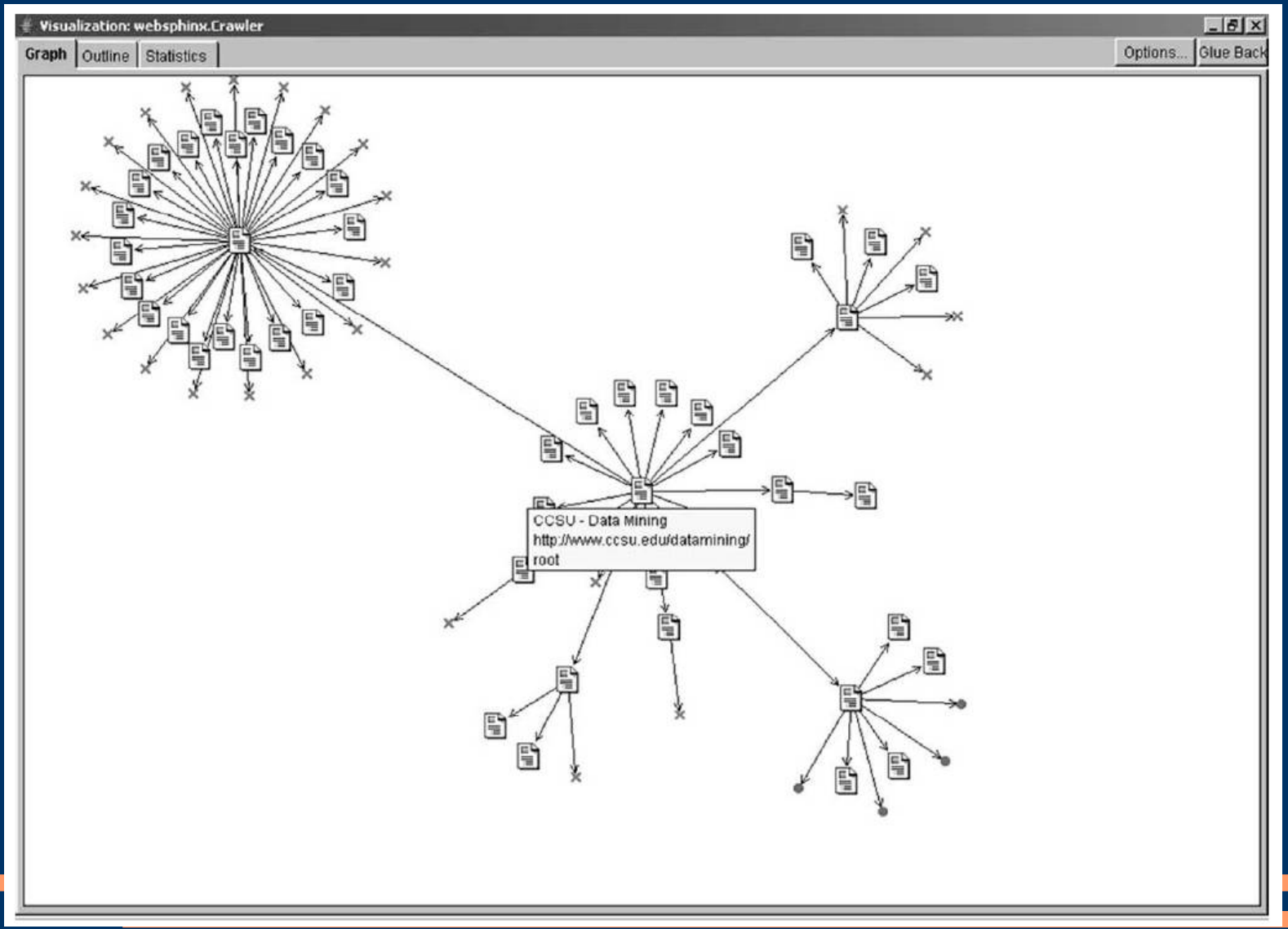
A_t – kolejka elementów do odwiedzenia,

s_t – aktualnie odwiedzany wierzchołek,

$N(s_t)$ – lista sąsiadów wierzchołka s_t.

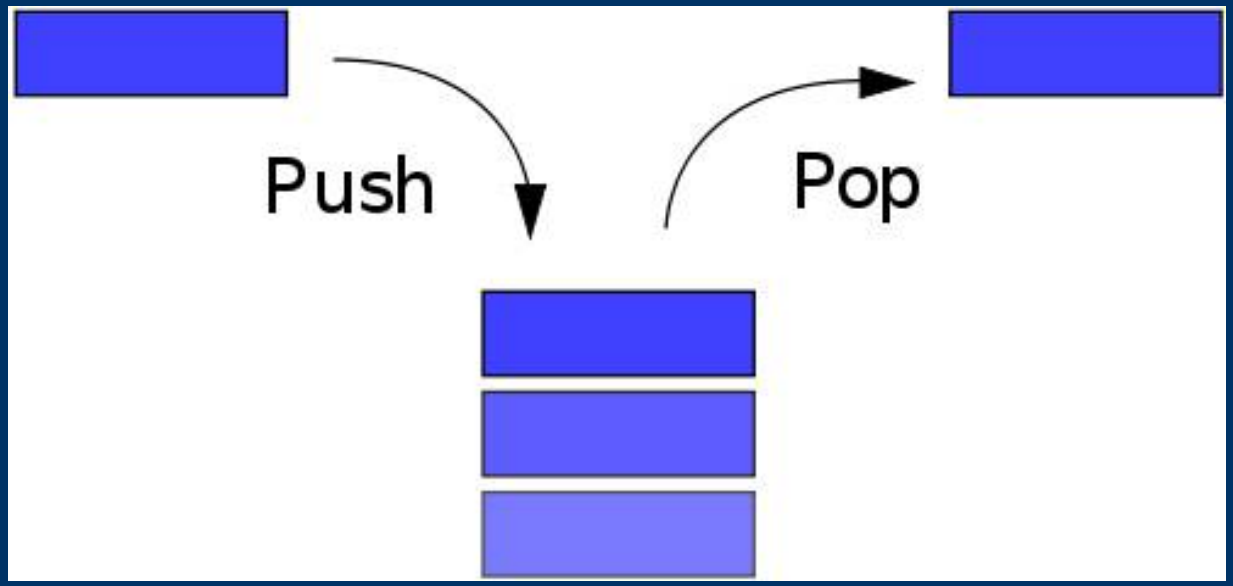
Przykład wszerz

Przykładowa reprezentacja grafu zbudowanego na bazie dokumentów ze strony http://www.ccsu.edu/datamining/



Przeszukiwanie w głąb

Metoda przeszukiwania w głąb oparta jest o strukturę LIFO.



Przeszukiwanie w głąb

Cechą przeszukiwania w głąb jest analiza każdej ścieżki do jej końca.

Wykorzystywane oznaczenia:

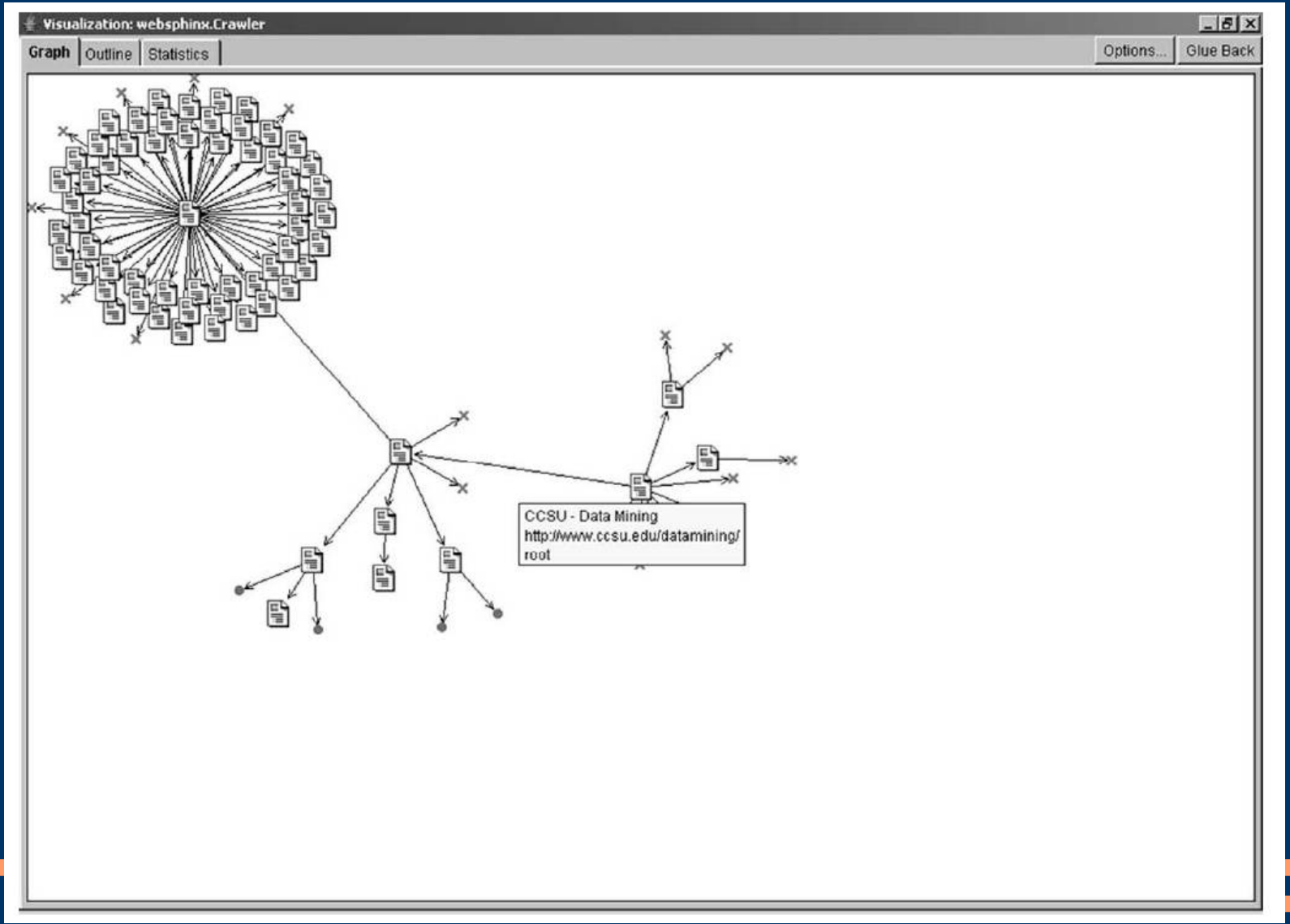
A_t – kolejka elementów do odwiedzenia,

s_t – aktualnie odwiedzany wierzchołek,

$N(s_t)$ – lista sąsiadów wierzchołka s_t.

Przykład w głąb

Przykładowa reprezentacja grafu zbudowanego na bazie dokumentów ze strony http://www.ccsu.edu/datamining/



Dziękuję za uwagę!