padkami harmonogramowania ponownych odwiedzin stron internetowych w przyrostowym modelu pozyskiwaniu dokumentów. Celem pozyskiwania dokumentów w badanym przypadku było jak najszybsze odnalezienie nowych informacji publikowanych na forach, bez niepotrzebnych pobrań powodujących obciążenie infrastruktury robota internetowego, jak również docelowych witryn.

1. Prace pokrewne

Prace, których zakres jest pokrewny do poruszanego tutaj, obejmują kilka obszarów problemowych: badania nad strukturą sieci i jej zmianami, metody wykrywania zmian, zagadnienia związane z projektowaniem robotów indeksujących działających w sposób przyrostowy oraz zagadnienia nawigowania pomiędzy stronami internetowymi, w tym poruszanie się pomiędzy stronami o zadanym typie (*focused crawling*).

1.1. Badanie zmienności stron internetowych

Naszą pracę opieramy na nowatorskim artykule napisanym przez Cho i Garcia-Molina [2000], w którym autorzy przedstawiają podobne badania ewolucji stron internetowych w celu zaprojektowania przyrostowego robota internetowego. Analiza, jakiej dokonali, obejmowała dużą kolekcję stron internetowych (ponad 500 000), a jej wynikiem była konkluzja, iż strony internetowe zmieniają się średnio co 10 dni (przy jednodniowej granulacji pomiaru), ze znaczącą różnicą pomiędzy stronami komercyjnymi i stronami ukierunkowanymi mniej biznesowo. Dodatkowo w artykule udowodniono empirycznie, że zmiany na stronach internetowych można opisać modelem matematycznym procesu Poissona, który jest używany do obliczania prawdopodobieństwa wystąpienia zdarzenia w sekwencji niezależnych zdarzeń losowych, charakteryzujących się stałą średnią częstością w czasie.

Ci sami autorzy kontynuowali swoją pracę nad udoskonaleniem matematycznego modelu do szacowania częstości zmian na stronach internetowych oraz nad optymalizacją czasu dostępu do zasobów przez przyrostowe roboty indeksujące [Cho i Garcia-Molina 2003]. Ich eksperymenty dowiodły, że używając odpowiednich estymatorów, robot indeksujący może w znaczny sposób podnieść swoją wydajność, co skutkuje większym odsetkiem wykrytych zmian.

Kolejne prace poszerzające wiedzę na temat charakterystyki zamian stron internetowych, skupiają się na optymalizacji robotów internetowych. W pracy przygotowanej przez Baeza-Yatesa i in. [2005] omówione zostały strategie dotyczące priorytetyzacji w procesie pozyskiwania dokumentów internetowych. Autorzy skupili się na strategiach biorących pod uwagę ważność stron internetowych (obliczaną głównie na podstawie PageRank). Na podstawie badań wywnioskowali, że użycie odpowiedniej strategii pozwala na wcześniejsze odnalezienie stron o lepszej jakości niż w przypadku stosowania strategii naiwnych, na przykład wykorzystujących przeszukiwanie wszerz.

Badania nad schematami zmian w obrębie stron internetowych zaprezentowali również Adar, Teevan i Dumais [2009]. Prowadzili oni analizy porównawcze o większej granulacji, wykorzystując pojedyncze elementy strukturalne stron, a nie dokumenty internetowe w całości. Po wybraniu próby 55 000 adresów URL prowadzono ich monitorowanie w odstępach godzinnych, co pozwoliło na dokładniejsze określenie czasu, w którym zachodziły zmiany na stronach. Badania pokazały, że częstość zmian na obserwowanych stronach była wyższa niż wskazywały na to poprzednie eksperymenty, przy czym znaczna część stron zmieniała się częściej niż co godzina. Wynikiem dokładnej analizy zawartości i struktury stron była identyfikacja zasadniczo stałych oraz wysoce zmiennych obszarów na każdej stronie. Korzystając z tej samej próby, autorzy wykazali również pewną zależność pomiędzy częstością odwiedzin strony internetowej przez użytkowników a schematami zmian na stronie [Adar, Teevan i Dumais 2009].

Nieco inne podejście zaprezentowali Saad i Gaŋcarski [2010], którzy starali się odnaleźć schematy w okresowych zmianach na stronach internetowych w celu umożliwienia robotowi internetowemu oszacowania momentu w czasie, kiedy kolejna zmiana pod danym adresem URL jest najbardziej prawdopodobna. W badaniu wykorzystano obserwację, że dla pewnych często aktualizowanych stron możliwe jest ustalenie stałej częstości zmian. To podejście zostało wykorzystane w zadaniach archiwizacji treści internetowych w celu poprawienia jakości wynikowego repozytorium dokumentów.

1.2. Metody wykrywania zmian

Metoda wykrywania zmian może być kluczowym elementem w pozyskiwaniu dokumentów z forów internetowych oraz w dostosowywaniu strategii ponownych odwiedzin stron internetowych w celu utrzymania aktualności lokalnej kolekcji dokumentów. Jednym z najprostszych rozwiązań jest użycie funkcji skrótu dokumentu do porównania kolejnych wersji danej strony. Metoda ta jest bardzo wydajna, jednakże wadą takiego podejścia jest to, że zarówno niewielkie, jak i duże zmiany traktowane są jednakowo. Fakt modyfikacji może być również stwierdzony na podstawie informacji zawartych w nagłówku HTTP Last-Modified, choć w przypadku stron generowanych dynamicznie nie zawsze można polegać na informacji tam zawartej. Wiele innych rozwiązań zostało zaproponowanych dla bardziej rozbudowanego opisu i wykrywania zmian.

Jednym z pierwszych rozwiązań zaprojektowanych do porównywania dokumentów HTML jest program HtmlDiff rozwijany przez AT&T Bell Laboratories [Douglis i Ball 1996; Douglis i in. 1998]. Program ten wykorzystuje algorytm zaproponowany przez Hirshberga do rozwiązania problemu znajdowania najdłuższego wspólnego podciągu [Hirschberg 1977]. Podczas szukania wspólnych ciągów tokenów HtmlDiff definiuje token jako znacznik kończący zdanie (np. <p>, <hr>, <li>) lub zdanie, które składa się z sekwencji słów oraz znaczników niekończących zdań (np. <b>, <a>) To rozwiązanie zostało później rozwinięte przez Chen i in. w implementacji programu

TopBlend [Farn Chen i in. 2000]. Charakteryzował się on wzrostem wydajności dzięki wykorzystaniu bardziej współczesnego algorytmu porównującego, zaproponowanego przez Jacobsona i Vo [1992].

Rocco, Buttler i Liu [Buttler, Rocco i Liu 2004; Rocco, Buttler i Liu 2003] przedstawili mechanizm zwany Page Digest do przechowywania i przetwarzania dokumentów internetowych. Page Digest wykazuje wiele zalet tradycyjnej funkcji skrótu dokumentu, jednocześnie pozwalając na odtworzenie oryginalnego dokumentu bez dodatkowego kosztu obliczeniowego. Rozwiązanie to oferuje przejrzystą separację zawartości strony internetowej od jej struktury, dostarczając bardziej użytecznej abstrakcji dokumentu internetowego niż reprezentacja tekstowa. Opisywany mechanizm pozwala również na bezpośrednie porównanie utworzonych skrótów dokumentów, jak również na zawężenie porównania do poszczególnych sekcji dokumentów. Eksperymenty zaprezentowane przez autorów potwierdzają, że wykrywanie zmian przy użyciu Page Digest może się odbywać z liniowym przyrostem potrzebnego czasu, wykazując 75-procentowe usprawnienie w porównaniu z innymi rozwiązaniami.

Yeh, Li i Yuan [2006] badali zagadnienie wykrywania zmian na stronach internetowych generowanych dynamicznie przy wykorzystaniu mechanizmu nadpisywania adresów (*URL rewriting*). Mechanizm ten stanowi wyzwanie dla rozwiązań z zakresu monitorowania stron internetowych, ponieważ adresy URL wewnątrz strony mogą zawierać identyfikator sesji użytkownika, który zmienia się przy każdym pobraniu strony, powodując, że dwie wersje tej samej strony internetowej, pozyskane w dwóch kolejnych sesjach, mogą prezentować się odmiennie. Autorzy zaproponowali zestaw wykazujących się dużą dokładnością algorytmów służących do odnajdowania podobieństw w kolejnych wersjach pozyskiwanych stron internetowych oraz do wykrywania zmian.

Z uwagi na to, że dokumenty HTML mogą być porównywane podobnie do wszystkich innych dokumentów tekstowych, Kwon, Lee i Kim [2006] poddali ewaluacji pięć istniejących miar podobieństwa, używając dokumentów internetowych z usuniętymi znacznikami HTML. Analiza dotyczyła: porównywania bajt po bajcie, cosinusowej miary TF-IDF, miary podobieństwa zbiorów słów, odległości edycyjnej oraz miary W-shingling. W pracy pokazano różnice pomiędzy wskazaniami poszczególnych miar w zależności od rodzaju zmian na analizowanych stronach.

Adar i in. [2009] koncentrowali swoją pracę wokół dynamiki zawartości stron internetowych. Odrzuciwszy funkcję skrótu dokumentu, prowadzili analizę zmian na stronach internetowych przy wykorzystaniu współczynnika Dice, mierzącego stopień pokrywania się tekstowej zawartości w porównywanych dokumentach. Autorzy następnie wzbogacili swój model o analizę zmian metodami działającymi na poziomie pojedynczych słów, a także o charakterystykę zmian strukturalnych na stronach internetowych. W artykule zaproponowano kilka algorytmów do opisu modyfikacji elementów drzewa DOM, jak również do określania trwałości poszczególnych bloków strukturalnych na stronie. W dalszych pracach autorzy podejmowali zagadnienie odnajdowania schematów w ewolucji stron internetowych i ich kolejnych zmianach, przy wykorzystaniu współczynnika Dice i tak zwanej "krzywej zmian" [Adar, Teevan i Dumais 2009].

Kilka bardziej współczesnych prac wykorzystuje cechy wizualne stron internetowych do wykrywania i pomiaru zmian. Saad i Gaŋcarski [2010] zaproponowali algorytm Vi-DIFF, który bierze pod uwagę segmentację wizualną strony internetowej, jak również zdefiniowane wskaźniki ważności dla różnorodnych zmian. Law i in. [2012] rozwinęli system do porównywania stron internetowych, wykorzystujący jednocześnie wizualne oraz strukturalne właściwości do opisu zmian na stronach.

1.3. Budowa przyrostowych robotów internetowych

Wspomniana wcześniej praca Cho i Garcia-Moliny na temat ewolucji stron internetowych [Cho i Garcia-Molina 2000] pozwoliła im na sformułowanie wytycznych związanych z projektowaniem przyrostowych robotów indeksujących. Znaczące rozbieżności w częstotliwości zmian na różnych stronach internetowych (wiele z nich nie zmieniło się przez cały 4-miesięczny okres eksperymentu) sugeruje, że pozyskiwanie poszczególnych stron powinno być podyktowane częstotliwością, z jaką kolejne strony się zmieniają, w celu ograniczenia zużycia zasobów na strony, które zmieniają się rzadziej. Dodatkowo proces Poissona, przyjmowany jako model zmian stron internetowych, sugeruje, że strategia przyrostowego pozyskiwania stron powinna być uzależniona od szacunków wskazujących, które strony z największym prawdopodobieństwem zmieniły się w danym okresie, zgodnie z modelem procesu i danymi historycznymi, i powinny być odwiedzone jako pierwsze.

Zaproponowano również wiele innych podejść do priorytetyzowania stron w celu ponownego pobrania, które nie ograniczają się jedynie do analizy częstości zmian stron internetowych zaobserwowanych w przeszłości. Jak pokazano w [Liu i in. 2011], miary ważności stron wykorzystujące informacje o grafowej strukturze sieci (jak stopień wejściowy wierzchołka lub PageRank) pozwalają na osiągnięcie dobrych rezultatów. Innymi czynnikami wpływającymi na porządek odwiedzania stron przez roboty indeksujące mogą być również: zakres tematyczny strony, struktura portalu, w którym strona się znajduje, a nawet informacje na temat zainteresowania użytkowników, pozyskiwane bezpośrednio z logów silników wyszukiwawczych.

1.4. Ukierunkowane pozyskiwanie stron

Istnieje kilka rozwiązań związanych z rozwojem metod i budową robotów internetowych, które skupiają się na pozyskiwaniu stron specyficznego rodzaju.

Pobieranie stron, ukierunkowane na kolekcjonowanie głównie nowych dokumentów pojawiających się w internecie, jest możliwe dzięki wykorzystaniu miary nowości zaproponowanej przez Toyodę i Kitsuregawę [2006]. Może być ona wykorzystana do odnajdowania pojawiających się nowych informacji, zgodnie z ustalonym zakresem tematycznym, np. zainteresowaniami użytkownika. Robot indeksujący może wykorzystać rozkład wartości miary nowości, obliczonej dla lokalnej kolekcji pobranych stron, do dostosowania strategii ponownych odwiedzin oraz odkrywania obszarów sieci, które z największym prawdopodobieństwem zawierają nowe strony.

Baeza-Yates i Castillo [2007] omawiają najpopularniejsze sposoby określania zakresu stron, które powinny być pozyskane z nieskończonej liczby stron, jakie mogą znajdować się w internecie. Można się ograniczyć do pobierania wyłącznie statycznych stron lub używać tylko jednego zestawu parametrów przekazywanych w adresie URL, można również ustalić maksymalne wartości graniczne określające liczbę stron (w całości lub dla poszczególnych domen) lub głębokość w strukturze witryny, których osiągnięcie kończy proces pobierania. Autorzy proponują różne modele opisujące losowe poruszanie się po uogólnionym modelu witryny i wskazują główne obszary, na których powinien się skupiać robot internetowy, z uwagi na to, jak głęboko w strukturę strony nawiguje przeciętny użytkownik.

Dużym zainteresowaniem badaczy cieszy się również pozyskiwanie dokumentów z forów internetowych. Cai i in. [2008] zbudowali inteligentnego robota działającego na forach internetowych, zwanego iRobot, który potrafi wybrać odpowiednie ścieżki przejścia przez strony różnego rodzaju wewnątrz forum dzięki analizie zawartości i struktury witryny. Robot próbuje zrekonstruować mapę witryny z użyciem wstępnego próbkowania stron z danego forum i grupowania ich zgodnie z charakterystyką rozmieszczenia treści. Mapa witryny pozwala aplikacji iRobot na wybór optymalnej kolejności pozyskiwania stron z uwzględnieniem zasobu informacyjnego na stronach, odrzucając duplikaty i strony zawierające błędy.

Yang i in. [2009] również zajmowali się tematem przyrostowego pozyskiwania stron internetowych z forów. Wykazali, że tradycyjne podejścia do pozyskiwania stron ogólnego przeznaczenia mogą być niewystarczające dla forów internetowych, które różnią się od zwykłych stron swoją budową i powiązaniami. Ustalenie strategii ponownych odwiedzin wykorzystujące wyłącznie częstość zmian strony może być mało wydajne, ponieważ w obrębie forum każda lista wątków lub lista postów (pojedynczy wątek) może być podzielona pomiędzy kilka stron, co powoduje, że jedna zmiana na liście może pociągnąć za sobą zmiany na wszystkich powiązanych stronach. A zatem zaproponowano strategię pozyskiwania stron wykorzystującą informacje o listach – po zrekonstruowaniu struktury powiązań posty z jednego wątku, które znalazły się na różnych stronach, zostają połączone w jedną listę, a następnie obliczany jest model regresji dla całego wątku.

FoCUS (Forum Crawler Under Supervision) został zaprezentowany przez Jiang, Yu i Lin. [2012]. Ich celem było zbudowanie robota internetowego działającego na forach, odwiedzającego jedynie relewantne dokumenty przy minimalizacji kosztów procesu ich pozyskiwania. Do zaprojektowania FoCUSa wykorzystano obserwację, że pomimo wykorzystywania różnych technologii do generowania struktury i zawartości stron internetowych, fora internetowe posiadają tak zwane ukryte ścieżki nawigacyjne, reprezentowane przez specyficzne adresy URL, które łączą strony początkowe ze stronami wątków. Autorzy zredukowali problem przechodzenia przez odpowiednie strony forów internetowych do problemu rozpoznawania typów adresów URL i umożliwili robotowi internetowemu uczenie się schematów przedstawianych jako wyrażenia regularne, opisujących ukryte ścieżki nawigacyjne. Pokazano, że przy wykorzystaniu

rozbudowanego klasyfikatora stron internetowych nawet niewielki zbiór uczący może być wystarczający do przeprowadzenia pozyskiwania stron na szeroką skalę.

1.5. Motywacja

Jak wynika z analizy literatury, istnieje wiele metod służących harmonogramowaniu procesu pozyskiwania stron internetowych. Jednakże celem tego artykułu jest przeanalizowanie podstawowych założeń, jakie są przyjmowane w tych modelach, i sprawdzenie, jakie schematy zmian dominują na stronach internetowych będących częścią forów. Przeprowadzone eksperymenty służą odnajdowaniu dalszych usprawnień metod harmonogramowania wykorzystujących zmiany na stronach internetowych lub wypracowaniu pewnego hybrydowego podejścia biorącego pod uwagę większą liczbę czynników, zmierzającego do poprawienia wydajności robotów indeksujących.

2. Eksperyment

2.1. Czym jest zmiana?

W celu analizy zebranych danych konieczne było szczegółowe określenie, jakie zdarzenie uważane jest za zmianę na stronie internetowej. Dane dostępne na potrzeby eksperymentu zawierały w sobie trzy podstawowe informacje o stanie każdej ze stron w kolejnych momentach w czasie:

- strona nie była dostępna pod danym adresem URL (oznaczone jako N/A),
- strona była dostępna pod danym adresem URL i była identyczna jak strona poprzednio pobrana spod tego adresu (oznaczone jako 0),
- strona była dostępna pod danym adresem URL, ale była inna (nawet jeżeli tylko o jeden bit) niż poprzednio pobrana strona spod tego adresu (oznaczone jako 1).

Ostatni przypadek obejmuje również sytuację, kiedy strona została pobrana po raz pierwszy lub nie była dostępna pod podanym adresem podczas poprzedniej próby pobrania, choć w pewnym momencie pobrano niepustą zawartość spod tego URL.

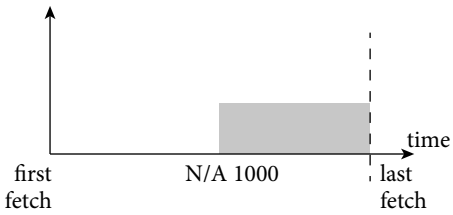
Z tego wynika, iż w praktyce może zaistnieć kilka sytuacji:

- N/A-N/A – strona niedostępna zarówno poprzednio, jak i w bieżącym pobraniu – brak zmiany,
- N/A-1 – strona niedostępna poprzednio, ale dostępna w bieżącym pobraniu,
- 1–1 – strona zmieniła się w poprzednim pobraniu i w bieżącym,
- 1-N/A – strona była zmieniona poprzednio, lecz jest niedostępna w bieżącym pobraniu,
- 1–0 – strona zmieniła się poprzednio, lecz w bieżącym pobraniu jest niezmieniona,
- 0–0 – strona nie zmieniła się ani poprzednio, ani w bieżącym pobraniu,
- 0-N/A – strona nie zmieniła się poprzednio, a w bieżącym pobraniu jest niedostępna,

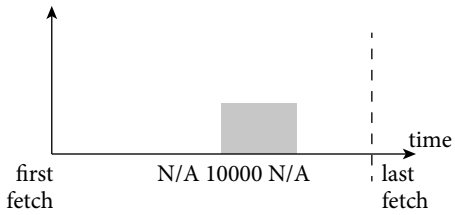
– 0–1 – strona nie zmieniła się poprzednio, a w bieżącym pobraniu nastąpiła zmiana.

Spośród wszystkich możliwości jedna (N/A-0) nie może wystąpić w prezentowanym zbiorze danych.

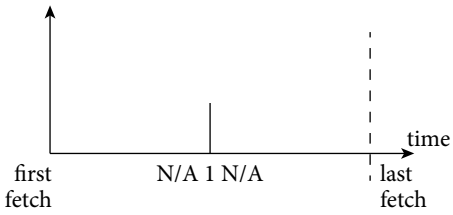
Taki sposób liczenia zmian jest umotywowany chęcią rozróżnienia stron na takie, które pojawiły się i nie zmieniły w późniejszym czasie, i takie, które pojawiły się i zniknęły zaraz potem lub po kilku pobraniach. Ta metoda daje następujące rezultaty w liczbie zaobserwowanych zmian dla prostych schematów dostępności stron:



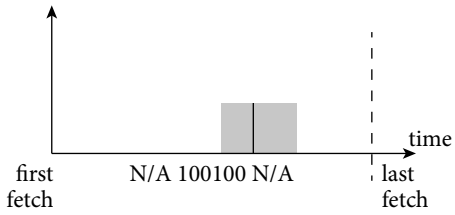
Rysunek 1. Strona była dostępna od pewnego momentu i nie zmieniła się do końca obserwacji. Liczba zmian: 1



Rysunek 2. Strona pojawiła się i stała się niedostępna, bez zaobserwowanych pośrednich zmian. Liczba zmian: 2



Rysunek 3. Strona pojawiła się jedynie na okres, w którym nastąpiło jedno pobranie. Liczba zmian: 2



Rysunek 4. Strona pojawiła się i po jednokrotnej zmianie pośredniej, stała się niedostępna. Liczba zmian: 3

2.2. Charakterystyka kolekcji danych

Do badania zostało wybranych 16 polskich forów internetowych, które były odwiedzane cyklicznie w trakcie eksperymentu. Po każdym odwiedzeniu forum tworzona była nowa migawka stanu forum składająca się z plików HTML reprezentujących strony internetowe należące do danego forum. Pliki były składowane w nieprzekształconej postaci w celu późniejszej analizy i porównywania.

W celu osiągnięcia dużej granulacji pobrań stron wybrane fora internetowe były odwiedzane co 2 godziny. Eksperyment trwał 23 dni (od 5 maja 2012 roku do 27 maja 2012 roku), co pozwoliło na zebranie 266 migawek dla każdego z forów. Ogólna liczba unikalnych adresów URL, które zostały pobrane w trakcie eksperymentu, to 27 958,

włączając w to strony, które zostały wykryte dopiero po pewnym czasie oraz strony, które w pewnym momencie stały się niedostępne. Aby określić liczbę zmian, zgodnie z algorytmem opisanym w sekcji 3.1, użyto danych z 649 705 udanych pobrań stron. Podsumowanie kolekcji danych pochodzących z eksperymentu jest przedstawione w tabeli 1.

Tabela 1. Podsumowanie kolekcji danych eksperymentalnych

Nazwa forum	Liczba pobrań	Liczba stron
forum.gazeta.pl	87 314	4 800
forum.o2.pl	15 009	86
forum.polskastrefa.eu	66 439	272
globtroter.pl	27 622	152
gumtree.pl	167 512	15 960
kafeteria.pl	35 868	3 267
oglaszamy24.pl	15 869	111
ogloszenia.adverts.pl	34 081	286
ojej.pl	22 362	119
olx.pl	45 013	581
owi.pl	18 629	1 604
pajeczyna.pl	43 067	181
top-ogloszenia.net	38 745	384
wrocek.pl	13 555	51
yahodeville.com	11 438	68
zdrowie-uroda.i-bazar.pl	7 182	36

Podobnie do pracy Cho i Garcia-Molina [2000] użyta została technika aktywnego monitorowania stron z zawężonym obszarem. Podczas pobierania stron robot internetowy odwiedza wybrane strony z pewną częstotliwością i zapisuje każdą wersję strony w celu dalszych analiz i wykrywania zmian. Aktywne monitorowanie może okazać się zbyt obciążające dla serwerów WWW dostarczających określone strony, zwłaszcza jeżeli pobieranie występuje zbyt często. Mimo wszystko został wybrany właśnie ten sposób monitorowania ze względu na możliwość precyzyjnego dopasowania parametrów pobierania stron, zakresu monitorowania, jak również otrzymania pewniejszych statystyk.

Podczas prezentowanego eksperymentu wybór stron był zależny od struktury danego forum. Dla każdej z witryn robot internetowy zaczynał pobieranie stron od predefiniowanej strony głównej i kontynuował pobieranie innych stron zgodnie z algorytmem przeszukiwania wszerz. Ta procedura była ograniczona do stron internetowych znajdujących się nie głębiej niż na piątym poziomie w strukturze nawigacyjnej forum. Stąd liczba pobieranych stron mogła się zmieniać podczas kolejnych

odwiedzin – niektóre strony mogły zostać dodane lub przeniesione do odpowiedniego poziomu w strukturze witryny, inne zaś mogły być usunięte lub przesunięte głębiej w strukturze.

Ograniczenie do odwiedzania stron leżących na maksymalnie piątym poziomie pozwoliło na ukończenie każdego procesu pobierania przed momentem, w którym należy rozpocząć kolejny proces. Oczywiście zakres monitorowania mógłby być poszerzony o kolejne poziomy w strukturze forów, lecz zważywszy na ograniczenia infrastruktury, pociągałoby to za sobą konieczność wydłużenia czasu pomiędzy kolejnymi pobraniami stron. Wstępne eksperymenty pokazały, że dla większych głębokości – 6 lub więcej – tak jak oczekiwano, w sposób znaczny wzrasta liczba pozyskiwanych stron podczas każdego procesu pobierania.

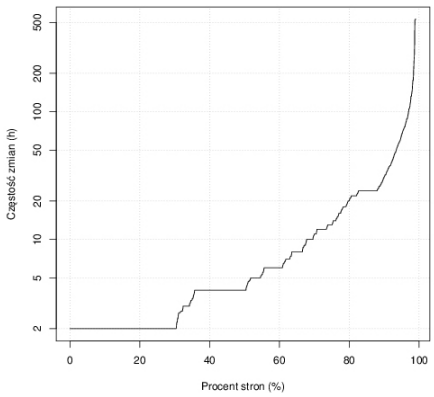
3. Rezultaty

Wyniki przeprowadzonych eksperymentów pokazują ciekawe własności stron internetowych należących do forów, zwłaszcza w kontekście poprzednich prac o zbliżonej tematyce. Ustalenie 2-godzinnej granulacji danych rzuciło nowe światło na częstość występowania zmian na stronach forów. Jak pokazano na rysunku 5, przeważająca liczba stron zmienia się bardzo często. Aż 30,42% monitorowanych stron zmieniało się średnio co 2 godziny, co oznacza, że za każdym razem, gdy dokonywane było pobranie strony, odnotowywana była zmiana – czy to w zawartości, czy w dostępności strony. Średnio co 4 godziny lub częściej zmieniało się 50,41% stron, a aż 87,96% zmieniało się średnio przynajmniej raz dziennie.

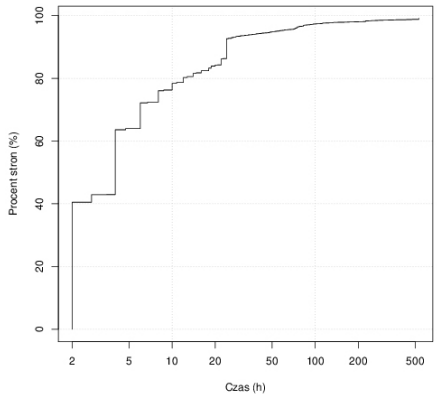
W celu określenia, jak szybko lokalna kolekcja dokumentów staje się nieaktualna, dla każdej ze stron zmierzono czas pomiędzy pierwszym pobraniem a pierwszą zmianą po pobraniu. Jak to przedstawia wykres 6, migawka stron pobranych z forum internetowego staje się nieaktualna w krótkim czasie – po 2 godzinach 40,49% stron się zmienia, po 4 godzinach – 63,6%, a potrzeba tylko 12 godzin, by 80,56% stron w kolekcji było przestarzałych.

Rysunek 7 i 8 przedstawiają wartości dotyczące odpowiednio dodanych i usuniętych stron, które zostały znalezione podczas eksperymentu, agregowane do jednodniowych okresów. Dane wskazują na bardzo wysoki współczynnik zmienności lokalnej kolekcji – każdego dnia przybywało średnio ponad 1000 nowych stron i podobna liczba stron stawała się niedostępna. Z uwagi na to, że średnia liczba stron w kolekcji każdego dnia wynosiła 3028, taka duża zmienność oznacza, iż każdego dnia 1/3 kolekcji była zastępowana przez nowe strony.

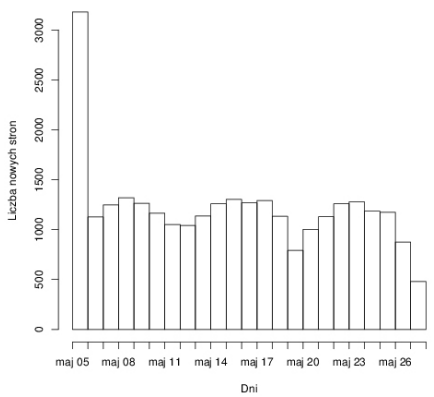
Jak można zauważyć, ciągi liczb, zarówno dotyczące nowych stron, jak i stron usuniętych, charakteryzują się tygodniową fluktuacją, ze wzrostem notowanym w środku tygodnia i spadkiem w okolicach weekendów.



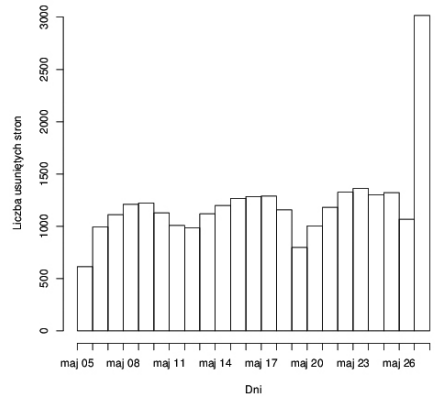
Rysunek 5. Procent stron o danej średniej częstości zmian (skala logarytmiczna)



Rysunek 6. Procent stron zmienionych po danym czasie (skala logarytmiczna)

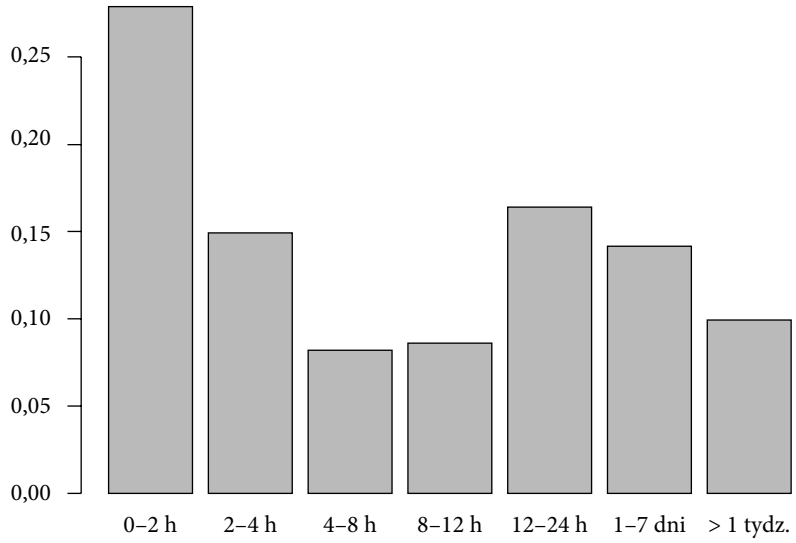


Rysunek 7. Liczba nowych adresów URL wykrytych w trakcie eksperymentu



Rysunek 8. Liczba adresów URL usuniętych w trakcie eksperymentu

Obliczono również czas dostępności każdej ze stron w trakcie eksperymentu jako czas pomiędzy pierwszym udanym pobraniem strony a ostatnim momentem, w którym strona była dostępna pod danym adresem URL, bez względu na zmiany lub okresową niedostępność strony. Dane zaprezentowane na rysunku 9 wskazują, że duża liczba stron ma relatywnie krótki czas dostępności – ponad 40% staje się niedostępnych w ciągu 4 godzin. Ponadto prawie 25% stron jest dostępnych przez okres dłuższy niż jeden dzień. Oczywiście przyjęte podeście do aktywnego pozyskiwania stron z narzuconym ograniczeniem mogło wpłynąć na przedstawione rezultaty w stopniu, którego nie można przewidzieć.



Wykres 9. Długość życia stron internetowych

Na końcu przeprowadzono analizę zgodności otrzymanych danych o zmienności stron internetowych z teoretycznymi wartościami rozkładu Poissona. Po słynnej publikacji Cho i Garcia-Molina [2000] popularnym podejściem w literaturze stało się wykorzystywanie rozkładu prawdopodobieństwa z procesu Poissona do prognozowania zmian na stronach internetowych.

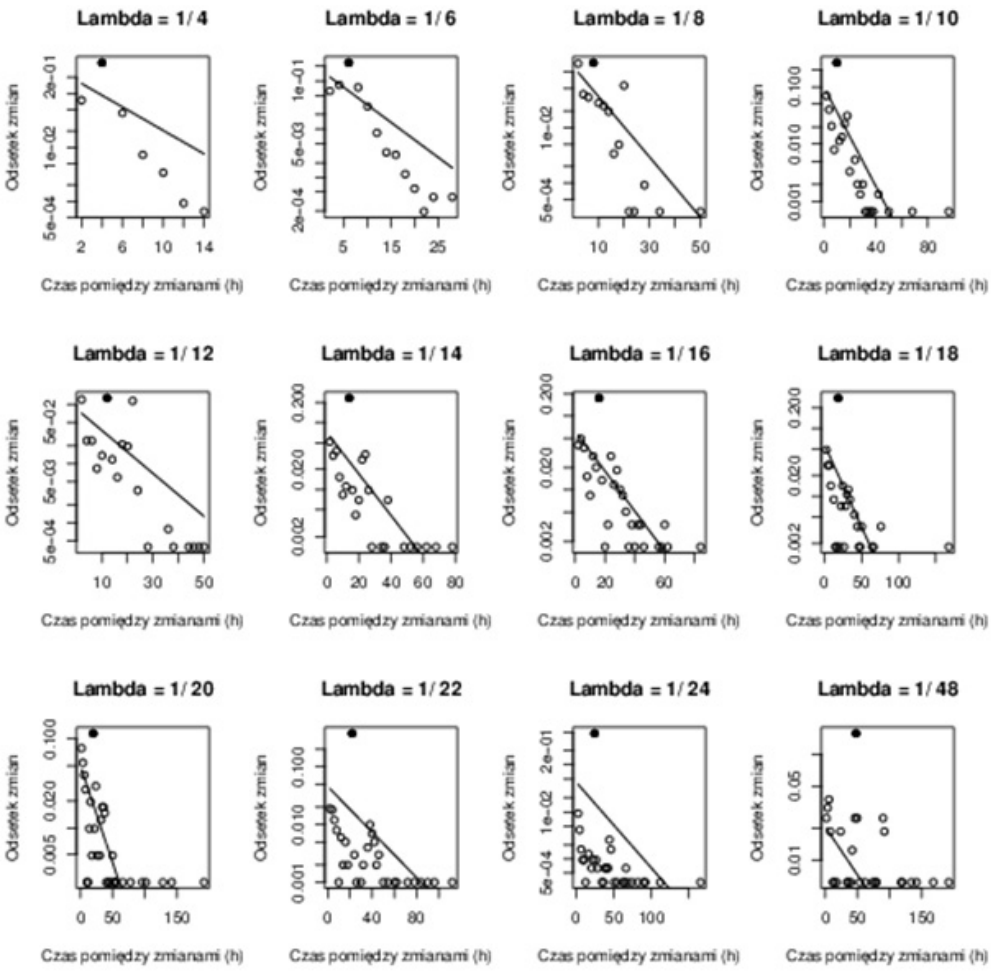
Rozkład prawdopodobieństwa czasu oczekiwania na wystąpienie zmiany na stronie, przy znanej średniej zmienności strony , dla nieujemnych wartości czasu jest określony wzorem:

$$f(t) = \lambda e^{\lambda t}.$$

Za pomocą tego wzoru można określić prawdopodobieństwo wystąpienia zmiany na stronie po określonym czasie, jeżeli znana jest średnia częstość zmian na tej stronie w przeszłości.

Wspomniani autorzy, korzystając z danych o dziennej granulacji, pokazali, że dla stron zmieniających się średnio np. co 10 lub co 20 dni dane empiryczne odpowiadają wartościom rozkładu Poissona. Na podstawie danych zebranych w prezentowanym eksperymencie przeprowadzono podobną analizę zmienności w krótszych okresach. Rysunek 10 przedstawia dane dotyczące czasu pomiędzy kolejnymi zmianami na monitorowanych stronach. Wybrane zostały strony o średniej częstości zmian od co 4 do co 24 godziny z odstępem dwugodzinnym oraz strony o średniej częstości zmian

co 48 godzin. Na rysunku umieszczono udziały okresów o odpowiedniej długości w ogólnej kolekcji okresów pomiędzy zmianami na stronach wraz z przebiegiem teoretycznych wartości rozkładu dla procesu Poissona dla odpowiednich parametrów. Jak można zauważyć, dla większej granulacji pomiarów oraz dla krótszych okresów średniej częstości zmian wartości empiryczne wykazują duże odstępstwa od rozkładu odpowiadającego procesowi Poissona. Szczególnie ciekawa jest obecność nieproporcjonalnie dużego odsetka obserwacji o wartościach równych średniej, co wskazuje na zmiany powtarzające się w nielosowy sposób.



Rysunek 10. Odsetek czasu pomiędzy kolejnymi zmianami zaobserwowanymi dla stron o danej średniej częstości zmian

Uważa się, że rozbieżności z rozkładem Poissona mogą być wynikiem specyfiki obszaru badania, tj. forów internetowych. Rozkład Poissona zakłada losowość występowania zmian na stronach przy zachowaniu stałej średniej. Jednakże fora internetowe tworzone są przez internautów, administratorów oraz mechanizmy automatyzujące w sposób bardziej usystematyzowany. Można się spodziewać, że moderowanie postów może być procesem ciągłym, natomiast usuwanie postów, które są przestarzałe, może się odbywać periodycznie w sposób wsadowy. Podobnie pojawianie się nowych wątków może być procesem zbliżonym do losowego, natomiast większej aktywności (a tym samym zmienności) można się spodziewać na stronach prezentujących nowsze posty, które są prezentowane w pewien określony sposób w strukturze forum, na przykład blisko strony głównej. Aby uchwycić wszystkie te elementy wpływające na zmienność stron w obrębie forów internetowych, potrzebne są dodatkowe badania, na szerszą skalę.

4. Podsumowanie wyników eksperymentu

Ponad 30% stron warto jest odwiedzać częściej niż co 2 godziny, a ponad 50% warto jest odwiedzać częściej niż co 4 godziny. Pozostaje pytanie o to, jak wiele z tych 30% stron jest zmienianych dynamicznie przez skrypty, bez udziału faktycznie nowych treści. Jednak, niezależnie od odpowiedzi na to pytanie, co wymaga dodatkowych badań o większej granulacji pobrań stron, wydaje się, że w celu stałego monitorowania forów internetowych ważne jest zidentyfikowanie często zmieniających się stron, aby można było pozyskać nowe treści natychmiast po tym, jak się pojawią, zwiększając tym samym współczynnik świeżości lokalnej kolekcji.

5. Implikacje dla budowy przyrostowych robotów internetowych

Główną motywacją do przeprowadzenia niniejszego eksperymentu było sprawdzenie, czy istniejące podejścia do przyrostowego pozyskiwania stron internetowych przez roboty indeksujące są dostosowane i dostatecznie sprawne, by monitorować fora internetowe. Pozyskane dane wskazują, że w przypadku monitorowania forów na szeroką skalę musi być wzięta pod uwagę duża zmienność stron internetowych wchodzących w skład forów. Jednym z głównych celów pozyskiwania stron jest zebranie nowych treści tak szybko, jak to tylko możliwe. Z względu na to, że większość stron w ramach forów internetowych zmienia się szybciej niż raz na 24 godziny, robot internetowy musi odwiedzać określone adresy URL dużo częściej. Mając na uwadze, że zwykle przepustowość łącza internetowego jest ograniczona, a roboty indeksujące muszą się stosować do polityki grzeczności, aby nie przeciążyć serwerów WWW, ważne jest, aby

pozyskiwanie stron było ukierunkowane na strony o większej częstotliwości zmian i zawierające bardziej relewantne treści. Jednak pewna losowość zmian na stronach internetowych sprawia, że planowanie kolejnych odwiedzin wyłącznie na podstawie częstotliwości zmian na stronie w przeszłości może się okazać niewystarczające. Inne aspekty, jakie mogłyby być wzięte pod uwagę, to na przykład dzień tygodnia lub struktura nawigacyjna forum. Połączenie informacji o częstości zmian na określonych stronach internetowych ze strukturą witryny w postaci grafu mogłoby wskazać elementy grafu, w których pojawiają się nowe treści. Jednocześnie byłaby możliwa optymalizacja polityki ponownych odwiedzin pozwalająca na pobieranie najważniejszych stron częściej niż co godzinę.

Bibliografia

Adar, E., Teevan, J. i Dumais, S.T., 2009, *Resonance on the Web: Web dynamics and Revisitation patterns*, w: *Proceedings of the 27th International Conference on Human Factors in Computing Systems*, CHI '09, ACM, New York, NY, USA, s. 1381–1390.

Adar, E., Teevan, J., Dumais, S.T. i Elsas, J.L., 2009, *The Web Changes Everything: Understanding the Dynamics of Web Content*, w: *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, ACM, New York, NY, USA, s. 282–291.

Baeza-Yates, R. i Castillo, C., 2007, *Crawling the Infinite Web*, J. Web Eng, vol. 6, no. 1, s. 49–72.

Baeza-Yates, R., Castillo, C., Marin, M. i Rodriguez, A., 2005, *Crawling a Country: Better Strategies than Breadth-first for Web Page Ordering*, w: *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, WWW '05, ACM, New York, NY, USA.

Ben Saad, M. i Gaŋcarski, S., 2011, *Archiving the Web Using Page Changes Patterns: a Case Study*, w: *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, JCDL '11, ACM, New York, NY, USA.

Buttler, D., Rocco, D. i Liu, L., 2004, *Efficient Web Change Monitoring with Page Digest*, w: *Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters*, WWW Alt. '04, ACM, New York, NY, USA, s. 476–477.

Cai, R., Yang, J.M., Lai, W., Wang, Y. i Zhang, L., 2008, *iRobot: An Intelligent Crawler for Web Forums*, w: *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, ACM, New York, NY, USA, s. 447–456.

Cho, J. i Garcia-Molina, H., 2000, *The Evolution of the Web and Implications for an Incremental Crawler*, w: *Proceedings of the 26th International Conference on Very Large Data Bases*, VLDB '00, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, s. 200–209.

Cho, J. i Garcia-Molina, H., 2003, *Estimating Frequency of Change*, ACM Trans. Internet Technol., vol. 3, no. 3, s. 256–290.

Douglis, F. i Ball, T., 1996, *Tracking and Viewing Changes on the Web*, w: *USENIX Technical Conference*, AT&T Bell Laboratories.

Douglis, F., Ball, T., Chen, Y.F. i Koutsofios, E., 1998, *The AT&T Internet Difference Engine: Tracking and Viewing Changes on the Web*, World Wide Web, vol. 1, s. 27–44.

Farn Chen, Y., Douglis, F., Huang, H. i phong Vo, K., 2000, *TopBlend: An Efficient Implementation of HtmlDiff in Java*, w: *World Conference on the WWW and Internet*, s. 88–94.

Hirschberg, D.S., 1977, *Algorithms for the Longest Common Subsequence Problem*, J. ACM, vol. 24, no. 4, s. 664–675.

Jacobson, G. i Vo, K.P., 1992, *Heaviest Increasing/Common Subsequence Problems*, w: A. Apostolico, M. Crochemore, Z. Galil i U. Manber (eds.), *Combinatorial Pattern Matching*, tom 644 z serii *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, s. 52–66..

Jiang, J., Yu, N. i Lin, C.Y., 2012, *FoCUS: Learning to Crawl Web Forums*, w: *Proceedings of the 21st International Conference Companion on World Wide Web*, WWW '12 Companion, ACM, New York, NY, USA, s. 33–42.

Kwon, S., Lee, S. i Kim, S., 2006, *Effective Criteria for Web Page Changes*, w: X. Zhou, J. Li, H. Shen, M. Kitsuregawa i Y. Zhang (eds.), *Frontiers of WWW Research and Development – APWeb 2006*, t. 3841 z serii *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg.

Law, M.T., Thome, N., Gançarski, S. i Cord, M., 2012, *Structural and Visual Comparisons for Web Page Archiving*, w: *Proceedings of the 2012 ACM Symposium on Document Engineering*, DocEng '12, ACM, New York, NY, USA, s. 117–120.

Liu, M., Cai, R., Zhang, M. i Zhang, L., 2011, *User Browsing Behavior-driven Web Crawling*, w: *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, ACM, New York, NY, USA, s. 87–92.

Rocco, D., Buttler, D. i Liu, L., 2003, *Page Digest for Large-scale Web Services*, w: *E-Commerce, 2003. CEC 2003. IEEE International Conference on*, s. 381–390.

Saad, M.B. i Gançarski, S., 2010, *Using Visual Pages Analysis for Optimizing Web Archiving*, w: *Proceedings of the 2010 EDBT/ICDT Workshops*, EDBT '10, ACM, New York, NY, USA, s. 43:1–43:7.

Toyoda, M. i Kitsuregawa, M., 2006, *What's Really New on the Web?: Identifying New Pages from a Series of Unstable Web Snapshots*, w: *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, ACM, New York, NY, USA, s. 233–241.

Yang, J.M., Cai, R., Wang, C., Huang, H., Zhang, L. i Ma, W.Y., 2009, *Incorporating Site-level Knowledge for Incremental Crawling of Web Forums: A List-wise Strategy*, w: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, ACM, New York, NY, USA, s. 1375–1384.

Yeh, P.J., Li, J.T. i Yuan, S.M., 2006, *Tracking the Changes of Dynamic Web Pages in the Existence of URL Rewriting*, w: *Proceedings of the Fifth Australasian Conference on Data Mining and Analystics – Volume 61*, AusDM '06, Australian Computer Society, Inc., Darlinghurst, Australia, Australia, s. 169–176.

VARIABILITY IN THE CONTENT OF INTERNET FORUMS

Abstract: In this article we present the results of a study conducted on a sample of Polish Web forums in order to investigate how these sites evolve over time. We analysed more than 27 900 Web pages from 16 sources at two hour intervals (4 256 data points) over 22 days of the experiment. The results can be the basis for improving Web crawler design, providing valuable insights into the nature of Web forums. It appears that the variability of Web forums content is significantly different from general-purpose Web sites, thus Web crawlers need to adjust their document extraction policies to deal with this kind of Web source.

Keywords: Internet forums, content variability, web crawlers.