Eksploracja zasobów internetowych
Kierunek Systemy teleinformatyczne, WE, II st., sem. II

Laboratorium nr 2

Celem zajęć jest ocena zastosowania wybranych narzędzi analitycznych do analizy wykorzystania zasobów wybranego serwera WWW.

Aby zaliczyć laboratorium **przygotuj sprawozdanie**, w którym umieść odpowiedzi do pytań oraz sformułowane wnioski w przeprowadzonej analizy. Sprawozdanie **prześlij** w systemie ilias.

Dane wejściowe: access.log wybranego serwera www

Jakie dane wykorzystałeś do ćwiczenia:……

Polecenia:

1. Dokonaj zgrubnej ceny danych zawartych w pliku access.log
   Jakiego rodzaju informacje możesz odczytać z zapisu poszczególnych zapytań do serwera?

Załóżmy, że będziemy analizować zachowania poszczególnych użytkowników serwera w oparciu i tzw. sesje.

2. W jaki sposób zidentyfikowałeś sesję?
   Zaprojektuj oraz zaimplementuje odpowiedni algorytm (skrypt, makro, itp.)
   Opisz zastosowaną procedurę/algorytm.
   Program/procedurę/algorytm dołącz do sprawozdania.
3. Jaki czas przyjąłeś do identyfikacji sesji?
   Jaki ten czas powinien generalnie być - czy maksymalnie długi, czy relatywnie krótki?
   Odpowiedź uzasadnij.
4. Jak najlepiej oszacować moment, w którym sesja się zakończyła?
5. Jak potraktowałeś żądania pojedyncze (jednokrotne)?
6. Zastosuj mechanizm odkrywania reguł asocjacyjnych.
   Jakie reguły asocjacyjne uzyskałeś? Ile jest tych reguł?
7. Co możesz powiedzieć o tych regułach w kontekście budowy serwisu oraz użytkowania serwisu?
8. Co możesz powiedzieć na temat topologii witryny na podstawie otrzymanej analizy asocjacyjnej?
9. Czy w Twoim zbiorze danych były wpisy robotów internetowych?
   Jakie to były wpisy?
   Co z nimi zrobiłeś i jak je zidentyfikowałeś?

Statystyki:

10. Ile różnych żądań zostało zarejestrowanych w analizowanym pliku logs?
11. Skąd były najczęstsze żądania?

Czy były to żądania od tej samej osoby?

12. Czy potrafisz zidentyfikować najczęściej odwiedzane strony podczas zidentyfikowanych sesji?
13. Czy te najczęściej odwiedzane strony były odwiedzane jako pierwsze w sesji?
14. Jak zmieniłaby się odpowiedź na pytanie nr 4, gdyby wydłużyć czas sesji?
15. Utwórz listę zasobów o największej liczbie żądań.
16. Utwórz listę katalogów o największej liczbie żądań.
17. Ile jest zasobów w poszczególnych katalogach?
18. Czy potrafisz coś powiedzieć o strukturze poszczególnych katalogów?
    Czy potrafiłbyś ocenić tę strukturę oraz zależności zasobów w tych strukturach?
19. Jak jest liczba żądań w sesji?
    Zbuduj histogram rozkładu.
20. Jak jest liczba żądań do poszczególnych katalogów?
21. Ile trwa typowa sesja?
    Zbuduj histogram czasów sesji.
22. Jaki jest średni czas trwania sesji?
23. Jaki jest rozkład średnich czasów przebywania na stronie?
24. Jaka jest zależność między długością sesji a liczbą żądań?
25. O ile średnio wydłuży się sesja, jeśli użytkownik zażądałby dodatkowej strony?
26. Jaki jest średni czas przebywania na danej stronie?
27. Jaki jest rozkład średniego czasu na stronę dla sesji składającej się z więcej niż jedno żądanie?
28. Jaka jest całkowita liczba wizyt?
29. Jaki jest średni czas wizyty?
    Zbuduj histogram czasów wizyt.
    Jak się mają te czasy do czasów sesji?
30. Jaki jest minimalny czas wizyty?
31. Jak jest całkowita liczba różnych użytkowników?
32. Jaka jest liczba użytkowników, jaka odwiedziła więcej niż jedną stronę, a ile takich którzy mieli więcej niż jedną sesję danego dnia?
33. Jaka jest średnia liczba wizyt dla użytkownika?
34. Sporządź statystykę wykorzystania serwera w poszczególnych dniach/tygodniach/etc.
35. Jaka jest liczba żądań/wywołać w sesji?