Eksploracja duĹĽych wolumenĂłw danych

Projekt – Sprawozdanie I

SformuĹ‚owanie problemu i przetwarzanie danych

ImiÄ™ Nazwisko ImiÄ™ Nazwisko

27 listopada 2012

1 SformuĹ‚owanie problemu

KrĂłtka informacja o wybranym konkursie: nazwa, strona internetowa, nagroda :), zakoĹ„czony.

1.1 KrĂłtki opis problemu

KrĂłtki i konkretny opis zadania konkursowego.
(maks. 1/2 strony)

1.2 Opis danych

Opis zbioru uczÄ…cego, walidujÄ…cego (jeĹĽeli jest) i testowego wraz z podstawowymi statystykami: rozmiar, liczba atrybutĂłw, liczba przykĹ‚adĂłw.

Warto przedstawiÄ‡ postaÄ‡ typowych przykĹ‚adĂłw uczÄ…cych oraz testowych.

NaleĹĽy rĂłwnieĹĽ opisaÄ‡ sposĂłb ostatecznej weryfikacji algorytmĂłw: trwajÄ…cy konkurs (wynik koĹ„cowy w konkursie; ocena na podstawie tablicy wynikĂłw, czyli na zbiorze walidujÄ…cym), moĹĽliwoĹ›Ä‡ weryfikacji po zakoĹ„czeniu konkursu na jego stronie, wewnÄ™trzny zbiĂłr testowy (rozmiar, jak zostaĹ‚, wyodrÄ™bniony, kiedy zostanie wykorzystany, itp.).

(maks. 3/4 strony)

1.3 Opis metody oceny rozwiÄ…zaĹ„

DokĹ‚adny i zwiÄ™zĹ‚y opis metody oceny rozwiÄ…zania (jeĹĽeli algorytm bÄ™dzie testowany wewnÄ™trznie, to naleĹĽy napisaÄ‡ odpowiednie oprogramowania oceniajÄ…ce rozwiÄ…zanie).

(maks. 1/2 strony)

2 Reprezentacja danych w pamiÄ™ci zewnÄ™trznej

KrÃłtki opis (bardzo zwiÄ™zÅ‚y i konkretny) z uwzglÄ™dnieniem:

- opisu reprezentacji z jej zaletami i wadami, rÃłwnieÅ¼ z perspektywy eksperymentÃłw, ktÃłre naleÅ¼y przeprowadziÄ‡,
- dyskusji na temat konfiguracji i administracji danego rozwiÄ…zania,
- dyskusji na temat czasu utworzenia i objÄ™toÅ›ci reprezentacji danych,
- dyskusji na temat napotkanych trudnoÅ›ci.

MoÅ¼na opisaÄ‡ wiÄ™cej niÅ¼ jedno rozwiÄ…zanie, w celach porÃłwnawczych. (maks. 2 strony)

3 Reprezentacja danych w pamiÄ™ci operacyjnej

KrÃłtki opis (bardzo zwiÄ™zÅ‚y i konkretny) z uwzglÄ™dnieniem:

- opisu reprezentacji z jej zaletami i wadami, rÃłwnieÅ¼ z perspektywy eksperymentÃłw, ktÃłre naleÅ¼y przeprowadziÄ‡,
- dyskusji na temat czasu wczytywania do pamiÄ™ci, dostÄ™pu do danych i objÄ™toÅ›ci danych,
- dyskusji na temat napotkanych trudnoÅ›ci.

MoÅ¼na opisaÄ‡ wiÄ™cej niÅ¼ jedno rozwiÄ…zanie, w celach porÃłwnawczych. (maks. 2 strony)

4 Eksperyment

4.0.1 Wyszukiwanie najbliÅ¼szych sÄ…siadÃłw

Wyszukiwanie najbliÅ¼szych sÄ…siadÃłw zostaÅ‚o przetestowane dla 9, 86, 862 i 8623 lekcji oraz x, xx, xxx i xxxx autorÃłw czyli dla odpowiednio 0,1%, 1%, 10% i 100%. Wyniki czasowe wraz z odchyleniem standardowym zebrano w tabelach.

W tym rozdziale naleÅ¼y umieÅ›ciÄ‡ wyniki eksperymentÃłw. NaleÅ¼y zamieÅ›ciÄ‡ krÃłtkie komentarze oraz tabele z wynikami (Tabela 2 jest przykÅ‚adem jak stworzyÄ‡ tabelÄ™ w LaTeX'u).

NaleÅ¼y wykonaÄ‡ nastÄ™pujÄ…ce zapytania:

1. Grupowanie,
2. Wyszukiwanie najbliÅ¼szych $k = 50$ sÄ…siadÃłw.

IloĹ›Ä‡ lekcji	Procent lekcji	Ĺšredni czas odpowiedzi	Odchylenie standardowe
9	0,1%	6754,6	734,93
86	1%	56039,4	4509,16
862	10%	6754,6	734,93
8623	100%	6754,6	734,93

Tablica 1: Wyniki eksperymentu dla wideo lekcji

To sÄ… kolumny	zawierajÄ… ce np.	czas odpowiedzi na zapytanie itp.
Opis i liczby:	99.99%	99.99%
Opis i liczby:	99.99%	99.99%
Opis i liczby:	99.99%	99.99%

Tablica 2: Wyniki eksperymentalne

Dla otrzymanych partycji danych (przez grupowanie i wyszukanie najbliĹĽszych sÄ… siadĂłw) naleĹĽy zastosowaÄ‡ dowolnÄ… funkcjÄ™ agregujÄ… cÄ… na dowolnych atrybutach. OczywiĹ›cie, najlepiej jest wybraÄ‡ atrybut decyzyjny (wyjĹ›ciowy).

Dla wszystkich eksperymentĂłw naleĹĽy podaÄ‡ czas wykonania.

NajbliĹĽszych sÄ… siadĂłw naleĹĽy wyszukaÄ‡ dla maksymalnie pierwszych 10 tys. przykĹ‚adĂłw, ale szukanie odbywa siÄ™ w caĹ‚ym zbiorze danych.

CaĹ‚y eksperyment moĹĽna powtĂłrzyÄ‡ parokrotnie i uĹ›redniÄ‡ wyniki (moĹĽna podaÄ‡ bĹ‚Ä… d standardowy). Przy wyznaczaniu czasĂłw nie naleĹĽy braÄ‡ pod uwagÄ™ czasu wyĹ›wietlania wynikĂłw.

W zaleĹĽnoĹ›ci od rodzaju problemu, proszÄ™, stosowaÄ‡ siÄ™ do poniĹĽej opisanych wytycznych. JeĹĽeli dany problem nie pasuje do ĹĽadnego wyĹĽej opisanego typu, to proszÄ™ o kontakt.

(maks. 4 strony (razem z tabelami)).

4.1 Tabelaryczne dane nominalne

W przypadku danych nominalnych naleĹĽy przeprowadziÄ‡ grupowanie:

- po wszystkich atrybutach (w przypadku duĹĽej liczby atrybutĂłw, grupowanie powinno dotyczyÄ‡ tylko pierwszych 20 atrybutĂłw),
- dla kaĹĽdego z atrybutu z osobna,
- oraz dla kaĹĽdej pary atrybutĂłw (maksymalnie dla pierwszych 100 atrybutĂłw).

Jako miarÄ™ podobieĹ„stwa przy wyszukiwaniu najbliĹĽszych sÄ… siadĂłw naleĹĽy zastosowaÄ‡ wspĂłĹ‚czynnik Jaccarda. Mierzy on podobieĹ„stwo

miÄ™dzy dwoma zbiorami i jest zdefiniowany jako iloraz mocy czÄ™Ś›ci wspÅ‚lnej zbiorÅ‚w i mocy sumy tych zbiorÅ‚w:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

gdzie A i B sÄ… zbiorami, ktÅ‚re odpowiadajÄ… obiektom opisanym przez atrybuty nominalne (proszÄ™ zwrÅ‚ciÄ‡ uwagÄ™, Ś‚e atrybut binarny czÄ™sto wskazuje, czy dana wartoÅ›Ä‡ wystÄ™puje, czy teÅ‚ nie).

Dla poniÅ‚szego przykÅ‚adu zbiÃ³r $A = \{\text{czerwony, duÅ‚y, kombi}\}$ odpowiada pierwszemu obiektowi, a zbiÃ³r $B = \{\text{niebieski, duÅ‚y, coupe, kabriolet}\}$ drugiemu. WspÃ³Å‚czynnik Jaccarda wynosi:

$$J(A,B) = \frac{1}{6}$$

kolor	rozmiar	kombi	sedan	coupe	kabriolet	limuzyna	minivan
czerwony	duÅ‚y	1	0	0	0	0	0
niebieski	duÅ‚y	0	0	0	1	1	0

MoÅ‚na teÅ‚ dokonaÄ‡ wczeÅ›niejszej binaryzacji wszystkich wielowartoÅ›ciowych atrybutÃ³w nominalnych i dalej operowaÄ‡ na atrybutach binarnych, ktÃ³re wskazujÄ… czy dany element wystÄ™puje w zbiorze.

4.2 Tabelaryczne dane numeryczne

W celu pogrupowania danych numerycznych dla kaÅ‚dego atrybutu naleÅ‚y przeprowadziÄ‡ dyskretyzacjÄ™ wartoÅ›ci numerycznych do 5 wartoÅ›ci. KaÅ‚da wartoÅ›Ä‡ odpowiada przedziaÅ‚owi wartoÅ›ci dla atrybutu numerycznego. NaleÅ‚y przeprowadziÄ‡ dyskretyzacjÄ™ wedÅ‚ug rÃ³wnej czÄ™stoÅ›ci, czyli otrzymane przedziaÅ‚y sÄ… mniej wiÄ™cej rÃ³wnoliczne.

Dla tak przekształconych atrybutÃ³w naleÅ‚y wykonaÄ‡ normalne grupowanie, zgodnie z z opisem w punkcie 4.1.

Przy poszukiwaniu najbliÅ‚szych sÄ…siadÃ³w naleÅ‚y siÄ™ posÅ‚uÅ‚yÄ‡ miarÄ… euklidesowÄ…. WczeÅ›niej naleÅ‚y jednak ustandaryzowaÄ‡ wartoÅ›ci atrybutÃ³w, tzn. od wartoÅ›ci kaÅ‚dego atrybutu odejmujemy Å›redniÄ… i dzielimy przez odchylenie standardowe (czyli przeprowadzamy zmienne o rÃ³Å‚Å‚nych jednostkach do zmiennych niemianowanych).

4.3 Tabelaryczne dane nominalne i numeryczne

W przypadku wymieszanych danych nominalnych i numerycznych dzielimy dane na dwie czÄ™Å›ci i wykonujemy eksperymenty osobno dla kaÅ‚dej czÄ™Å›ci zgodnie z opisem w dwÃ³ch powyÅ‚szych punktach.

4.4 Dane tekstowe

Dane tekstowe naleĹĽy zamieniÄ‡ na dane nominalne, gdzie kaĹĽdy atrybut odpowiada jednemu sĹ‚owu. Eksperyment naleĹĽy wykonaÄ‡ dla 100 najczÄ™Ĺ›ciej wystÄ™pujÄ…cych sĹ‚Ăłw. MoĹĽna (nie jest konieczne) wykonaÄ‡ odpowiednie przeksztaĹ‚cenia danych tekstowych, takie jak stematyzacja lub lematyzacja.

Dla tak przeksztaĹ‚conych danych naleĹĽy postÄ™powaÄ‡ zgodnie z opisem w punkcie 4.1.

4.5 Dane macierzowe

W niektĂłrych konkursach dane majÄ… postaÄ‡ zaleĹĽnoĹ›ci pomiÄ™dzy dwoma typami obiektĂłw, np. uĹĽytkownikami i produktami. W takim przypadku naleĹĽy wykonaÄ‡ grupowanie dla kaĹĽdego typu obiektĂłw osobno. JeĹĽeli jedynÄ… informacjÄ™ o obiekcie jest jego identyfikator, to grupowanie dotyczyÄ‡ bÄ™dzie tylko jednego atrybutu. W innym przypadku naleĹĽy stosowaÄ‡ siÄ™ do opisĂłw w powyĹĽszych punktach.

Wyszukiwanie najbliĹĽszych sÄ…siadĂłw w tym przypadku powinno polegaÄ‡ na wyszukiwaniu podobnych uĹĽytkownikĂłw ze wzglÄ™du na wybrane produkty, oraz wyszukiwaniu podobnych produktĂłw ze wzglÄ™du na "wybierajÄ…cych" uĹĽytkownikĂłw. W obydwĂłch przypadkach naleĹĽy wykorzystaÄ‡ wspĂłĹ‚Ĺ›czynnik Jaccarda w celu obliczenia podobieĹ„stwa.

5 Podsumowanie

Na koĹ„cu jest zawsze miejsce na krĂłtkie podsumowanie (maks. 1/2 strony). **CaĹ‚oĹ›Ä‡ raportu nie moĹĽe przekraczaÄ‡ 8 stron.**