Марина Георгиева
25767, ИИОЗ
4 September 2018

Домашно 1

Препоръчваща система базирана на съдържание

1. Описание на решението

Реализирана е препоръчваща система за препоръчване на нови документи на основа на прегледани от потребител документи формиращи неговия профил. Използван е корпусът с документи 20newsgroups. За индексиране и търсене е изпоплзван Apache Lucene.

Първата стъпка от решението е формирането на профил на потребителя. Избираме произволно няколко различни категории от 20-те възможни в корпуса. Взимаме всички новини от тях и ги разбъркваме. След това избираме N на брой новини от избраните категории, като броят на разгледаните от потребителя новини в профила му е предварително фиксиран (в случая използвам N=15).

Всички останали документи, които не се включват в потребителския профил, индексираме чрез Lucene. За всяка от прегледаните от потребителя новини в профила му търсим новините, с които най-много си прилича. За целта използваме MoreLikeThis заявката, предоставена от Lucene. Органинчаваме се до топ 5 резултатите за всяка новина от потребителския профил. Така ако сме имали N=15 новини в потребителския профил, след намирането на най-подобните 5 за всяка, може да получим общо 75 новини, от които ще избираме препоръките. Възможно е някои новини да се повтарят в тези 75, тоест една от непрочетените новини да е избрана като най-близка на две или повече новини от потребителския профил. Затова се налага филтрация преди да можем да направим предсказване за препоръките. След като сме филтрирали повтарящите се новини, ако има такива, сортираме новините по оценката от Lucene за тях (използваме вградената на Lucene score функция). Генерираме препоръките, като избираме топ 10 от тези новини спрямо техния score.

2. Резултати

Примерен резултат от препоръчващата система при даден потребителски профил:

Прегледани от потребителя новини:

```
1. Category: talk.religion.misc - 83438
2. Category: alt.atheism - 54133
3. Category: alt.atheism - 53648
```

```
4. Category: alt.atheism - 53106
5. Category: sci.crypt - 15977
6. Category: alt.atheism - 51204
7. Category: talk.religion.misc - 83993
8. Category: talk.religion.misc - 82781
9. Category: sci.crypt - 15769
10. Category: alt.atheism - 53598
11. Category: alt.atheism - 54198
12. Category: talk.religion.misc - 84214
13. Category: talk.religion.misc - 83752
14. Category: sci.crypt - 15869
15. Category: alt.atheism - 53421
```

Препоръки, изведени от системата:

```
1. Category: talk.religion.misc - 83989
2. Category: alt.atheism - 54184
3. Category: sci.crypt - 15576
4. Category: alt.atheism - 53409
5. Category: alt.atheism - 53495
6. Category: alt.atheism - 53412
7. Category: talk.religion.misc - 84075
8. Category: talk.religion.misc - 83998
9. Category: alt.atheism - 53653
10. Category: alt.atheism - 53582
```

В случая предложените препоръки изглеждат добри. В профила на потребителя преобладават новини от категорията Атеизъм и съответно в препоръките повече от половината предложени новини са от тази категория. Значи системата правилно е хванала, че тази тема е най-интересна за потребителя. Предложила му е и 3 новини от категорията Религия, тъй като тя е втората по-предпочитана тема от потребителя. Има и една новина свързана с криптиране, тъй като все пак е чел и такива новини, макар и по-малко. Вижда се, че в общи линии системата спазва съотношението на прочетените новини по категории, когато прави препоръки. Предлага най-много новини от най-четената категория и по-малко от другите. Също така виждаме, че не се е появила някаква новина, която е от съвсем различна категория от тези, които интересуват потребителя. Това показва, че системата наистина търси най-близките новини до тези в профила на потребителя.

3. Възможни подобрения

Ако имахме информация за това колко пъти е четена дадена новина от всички потребители на системата, можеше да вземем предвид това при генерирането на препоръките и да разглеждаме с по-голяма тежест новини, които са четени много пъти от други потребители и са подобни на новините в потребителския профил. Също така ако имаме информация за поведението на повече потребители в системата, може да открием потребители, които са подобни на нашия, и да предлагаме новини, които те са чели, а нашият потрбител - не.