

2014



Софийски университет „Св. Климент Охридски“

Факултет по математика и информатика

Катедра "Информационни технологии"

Система за препоръчване на съдържание в социалната мрежа

ДИСЕРТАЦИЯ

на Милен Митков Чечев

за присъждане на образователна и научна степен "доктор"

в професионално направление

4.6 „Информатика и компютърни науки“

научна специалност 01.01.12. „Информатика“

(Изкуствен интелект: Препоръчващи системи)

Научен ръководител:

доц. д-р Иван Койчев

юни 2014 г.

гр. София

Съдържание

Индекс на фигурите.....	6
Индекс на таблиците	7
Благодарности.....	8
Увод	9
Мотивация.....	10
Цел на дисертационния труд.....	11
Задачи на дисертационния труд	11
Структура на дисертационния труд	11
Глава 1. Обзор	13
1.1. Препоръчващи системи базирани на съдържание	14
1.1.1. Векторно-пространствен модел	16
1.1.2. Наивен бейсов класификатор	16
1.1.3. Подход базиран на алгоритъма на Рочио	17
1.2. Препоръчващи системи, базирани на сътрудничество	17
1.2.1. Сътрудничество използващо най-близкия съсед	18
1.2.1.1. Най-близкият съсед на база потребители	19
1.2.1.2. Най-близкият съсед на база продукт.....	20
1.2.2. Изграждане на модел с латентни свойства	20
1.2.3. Сравнение между препоръките базирани на съдържание и на сътрудничество.....	22
1.3. Препоръчващи системи, използващи демографски фактори	23
1.4. Препоръчващи системи, използващи знания	24
1.5. Препоръчващи системи, базирани на общности.....	24
1.6. Хибридни препоръчващи системи.....	26
1.6.1. Хибридна система с монолитен дизайн.....	26

1.6.2.	Хибридна система с паралелен дизайн	27
1.6.3.	Хибридна система с конвейерен дизайн	27
1.7.	Оценяване на препоръчващи системи	28
1.7.1.	Средна абсолютна грешка	28
1.7.2.	Средна Квадратична Грешка	28
1.7.3.	Метрики за точност от областта извличане на информация.....	29
1.7.4.	Средно Аритметична точност	30
1.7.5.	РОК-крива.....	30
1.7.6.	Среден Реципрочен Ранк	31
1.7.7.	Нормализирана обезценена кумулативна печалба.....	31
1.7.8.	Онлайн оценяване.....	32
1.7.9.	Свойства на препоръките.....	32
1.7.9.1.	Покритие.....	33
1.7.9.2.	Доверие.....	33
1.7.9.3.	Новост.....	34
1.7.9.4.	Изненадващи препоръки	34
1.7.9.5.	Разнообразност	34
1.7.9.6.	Полезност	35
1.7.9.7.	Устойчивост	35
1.7.9.8.	Адаптивност.....	35
1.7.9.9.	Разширяемост	36
Глава 2.	Изследване и анализ на проблема.....	37
2.1.	Изследване на необходимостта на потребителите.....	37
2.2.	Свързани изследвания	41
2.3.	Изводи.....	42
Глава 3.	Препоръчване в социалната мрежа	43
3.1.	Препоръки базирани на съдържание.....	43

3.2. Препоръки базирани на сътрудничество	44
3.3. Препоръки базирани на доверие.....	47
3.4. Отчитане на времевият фактор.....	48
3.5. Дискусия	50
Глава 4. Система за препоръчване в социалната мрежа	53
4.1. Функционалност.....	53
4.1.1. Четец за новини.....	53
4.1.2. Търсеща машина за новини	56
4.1.3. Възможност за социална активност в приложението	57
4.1.4. Възможност за персонално отбелязване на интересни и безинтересни новини.	57
4.1.5. Изграждане на автоматичен потребителски профил	57
4.1.6. Възможност за фина ръчна настройка на потребителския профил	59
4.2. Архитектура на системата.....	60
4.3. Потребителски интерфейс.....	61
Глава 5. Софтуерни инструменти и реализация	65
5.1. Платформи и софтуерни инструменти.....	65
5.1.1. Lens Kit.....	65
5.1.2. Apache Mahout.....	65
5.2. Достъп до данни в социалните мрежи.....	66
5.2.1. Facebook API.....	66
5.2.2. Twitter API.....	69
5.3. Прототип	70
Глава 6. Оценяване	72
6.1. Събиране на данни	72
6.2. Офлайн оценяване на системата.....	73
6.2.1. Изследване за правилният брой резултати за препоръчване	73

6.2.2. Оценяване на подредбата на списъка с препоръки.....	77
6.3. Онлайн оценяване на системата	79
Заключение.....	82
Перспективи за бъдещо развитие	83
Научни приноси.....	83
Научно-приложни приноси	83
Приложни приноси	84
Публикации.....	84
Доклади.....	85
Декларация за оригиналност	85
Библиография	86

Индекс на фигурите

Фигура 1. Архитектура на система за препоръчване на информация базирана на съдържание.....	15
Фигура 2. Таблица с оценки потребител-продукт.....	18
Фигура 3. Архитектура на хибридна система с паралелен дизайн.....	27
Фигура 4. Архитектура на хибридна система с конвейерен дизайн.	28
Фигура 5. Примерна РОК-крива	31
Фигура 6. Демографска информация за участниците в анкетата	37
Фигура 7. Време прекарвано във Фейсбук от участниците в анкетата.....	38
Фигура 8. Време, отделяно от участниците в анкетата, за различните дейности във Фейсбук.	38
Фигура 9. Избор на нова функционалност от Фейсбук потребителите.....	39
Фигура 10. Избор на опции за потока с новини във Фейсбук.	40
Фигура 11. Таблица на харесването на социалните обекти от потребителите.....	45
Фигура 12. Времева промяна на оценката с различна степен на социален интерес.	50
Фигура 13. Изгледи във Фейсбук: Топ-новини и най-скорошни новини.....	54
Фигура 14. Показване на реклама в Google.....	55
Фигура 15. Показване на реклама във Фейсбук.....	55
Фигура 16. Представяне на предпочитанията към потребители от скрипта на Sreedharan.....	58
Фигура 17. Явно събиране на предпочитания за новини във Фейсбук.	59
Фигура 18. Архитектура на Фейсбук приложение за препоръчване на информация	61
Фигура 19. Изглед от приложението за нов потребител.	62
Фигура 20. Основен изглед от приложението.	63
Фигура 21. Изглед от страницата за потребителски профил.....	63
Фигура 22. Резултат от извикване на Facebook Graph Api.....	69
Фигура 23. Точност и откриване с препоръки базирани на съдържание.....	74
Фигура 24. Точност и откриване на препоръки базирани на сътрудничество.	75
Фигура 25. Точност и откриване на препоръки базирани на доверие.....	76
Фигура 26. Оценяване на точността на хибридният подход.	77

Индекс на таблиците

Таблица 1.Таблица с различните видове обекти предоставяни от програмният интерфейс на Фейсбук	68
Таблица 2. Статистики за събраните данни.....	73
Таблица 3. Сравнение на подредбата на резултатите с различните подходи.....	78
Таблица 4. Резултати от потребителска анкета за реализираното приложение. ...	80

Благодарности

Благодаря за подкрепата и съвместната работа през годините на научния ми ръководител доц.Иван Койчев, колегите ми от Факултета по Математика и Информатика, Онтотекст, Амазон, на студентите активно участващи в новосъздаденият курс по препоръчващи системи и на дипломантите, които съм ръководел в това направление.

Увод

Препоръчващите системи започват развитието си в средата на 90-те години на миналият век (Resnick, Iacovou, Suchak, Bergstrom, & Riedl, 1994). Задачата, която се решава е за множество от потребители и множество от предлагани продукти, да се предскаже кои продукти ще бъдат интересни за различните потребители. Решението на тази задача дава възможността електронен магазин да разбере предпочитанията на потребителите си и да ги насочи към правилните продукти оптимизирайки по този начин работата на потребителя в магазина. Изследванията показват, че интернет магазините, които предлагат препоръки на потребителите си, успяват да покажат на потребителя по-широк асортимент от интересни за него продукти и да увеличат приходите си. Голямата възвращаемост на инвестицията определя силната им комерсиална насоченост и големият интерес към тях (Schafer, Konstan, & Riedl, 1999).

С развитието на интернет и навлизането на Web 2.0 (O'Reilly, 2007) предизвикателствата пред препоръчващите системи се променят. Потребителите започват да имат по-широка активност в мрежата, увеличава се желанието за споделяне на информация и за отбелязване (tagging) на различни мултимедийни ресурси. Развиват се социалните мрежи, в които потребителите имат възможността за електронна комуникация и споделяне на ресурси. Според статистиките в края на 2013, интернет потребителите са около 2,7 милиарда (UN's Millenium Development Goals Report, 2013), а в най-голямата социалната мрежа - Фейсбук активните потребители вече са повече от 1.3 милиарда, като само за последната една година са се увеличили с 22% (Facebook Statistics, 2014).

Атрактивността на социалните мрежи и увеличаването на информационните обекти, публикувани в тях, довежда до затрупване на потребителите с информация. Естественото решение на този проблем е в използването на препоръчващи системи, които да правят автоматично отсяване на интересните за потребителите обекти. Тези препоръчваща система могат да се използват в множество различни области в социалните мрежи – препоръчване на различни приятелства на потребителя за хора, които може да познава, препоръчване на различни информационни източници, , препоръчване на

публикувани новини, препоръчване на обекти публикувани от приятелите на потребителя, препоръчване на игри, препоръчване на активности и др.

Препоръчването на различни информационни обекти в социалната мрежа цели увеличаване на лоялността към социалната мрежа и поддържане на атрактивността и. Допълнително с разбиране на потребителските предпочитания, тази информация се използва от социалната мрежа за генериране на приходи, чрез персонализирани реклами и за различни социални експерименти (Booth, 2014).

Мотивация

Социалните мрежи са инструмент, който хората в повечето случаи използват за запълване на свободното си време за забавление, поддържане на онлайн приятелства и споделяне на информация. Голямото им разпространение ги прави много милиарден бизнес, като най-голямата социална мрежа Фейсбук генерира 7.87 милиарда долара приходи за 2013г (Rushe, 2014). Основните приходи на социалната мрежа идват от предлаганата от нея персонализирана реклама, като основните фактори за запазване и увеличаване на приходите им са предоставяне на услуга, удобна за потребителите, и по-добро разбиране на интересите на потребителите.

Според направено проучване, като част от дисертацията, основното време прекарано от потребителите в социалната мрежа е за разглеждане на потока от новини. Подредбата на новините е от изключителна важност за задоволяване на потребителите, което пък определя тяхната лоялност към мрежата. Новините, които участват в информационния поток са много, което предполага използването на препоръчваща система, която да оценява интереса на потребителя към новините и да генерира подредбата им спрямо това.

Проблемът за препоръчване на съдържание в социалните мрежи е сред най-актуалните проблеми в сферата на препоръчващите системи, като всяко подобрене на съществуващите системи би довело до увеличаване на доверието и лоялността на потребителите, и оттам увеличаване на приходите на социалната мрежа.

Обект на изследването е ежедневно взаимодействие на потребителите в социалната мрежа с техните приятели и с информацията споделена в социалната мрежа. Изследва се възможността да се подреди информационният поток на потребителя в ред, отговарящ на предпочитанията му, като се увеличи доверието на потребителя в предоставената му последователност от новини.

Предмет на изследването е създаване на система за препоръчване на информация на потребителите, използваща данните за потребителя от социалната мрежа и предлагаща алтернативен изглед на социалния му поток.

Изследователският въпрос, който е поставен в настоящият труд е: „Как да се подреди информационният поток за потребителите на социалните мрежи?”

Цел на дисертационния труд

Настоящия дисертационен труд има за цел да изследва нуждата на потребителите на социалните мрежи от система за препоръчване на информация, да анализира възможностите за изграждане и внедряване и да намери подходящ подход за сортиране на социалният поток от информация.

Задачи на дисертационния труд

Задачите на дисертационния труд са следните:

- Обзор на съществуващите подходи за препоръчване
- Изследване на нуждите на потребителите от нов тип препоръчваща система за информация от социалната мрежа.
- Разработване на подход за препоръчване на информация, базиран на резултатите от предишната точка.
- Изграждане на прототип
- Оценяване на ефективността на предложения подход.

Структура на дисертационния труд

Настоящият **увод** обосновава актуалността на проблема, дефинира обекта и предмета на дисертационния труд, представя целта и задачите на настоящата разработка. В допълнение представя структурата на работата.

Първа глава е отделена за обзор на проблемната област. В нея са показани различните видове препоръчващи системи, техните свойства и начини за оценяване.

Във **втора глава** се представя анализ на проблема. Представя се изследване с потребители от социалната мрежа, в което те определят от какво имат необходимост при работата със социалната мрежа.

Трета глава представя разработените препоръчващи подходи приложени за решаване на проблема за препоръчване на информация в социалната мрежа. Представени са: нов подход за пресмятане на доверието на потребител в неговите приятели и подход за остаряване на новини.

В **четвърта глава** се представя разработеното като част от дисертацията приложение за социалната мрежа Фейсбук. Дефинира се функционалността, която е необходима за реалното му използване. Изгражда се архитектурата на приложението и се дават детайли от реализацията.

В **пета глава** се дават детайли относно реализацията на приложението, известни софтуерни платформи за препоръчващи системи и програмният интерфейс към избрани социални мрежи.

В **шеста глава** са представени направените експерименти с реални потребители от социалната мрежа за оценяване на реализиране на препоръчващите подходи и реализираният прототип.

В края на дисертационния труд са разположени заключение, списък с приносите, публикациите и списък с използваната литература

Глава 1. Обзор

Изследванията в областта на препоръчващите системи започват през 90-те години миналият век (Resnick & Varian, 1997) (Schafer, Konstan, & Riedl, 1999). Основните подходи, които се разглеждат, са препоръки базирани на съдържание и препоръки използващи сътрудничество (Breese, Heckerman, & Kadie, 1998) (Resnick, Iacovou, Suchak, Bergstrom, & Riedl, 1994). Препоръките използващи сътрудничество са широко изследван подход и в началото на 21 век. Един от начините за стимулиране на изследователската работа в това направление е чрез онлайн състезания върху реални данни. Наградата на Нетфликс (Bennett, 2007) е най-голямото състезание в областта на препоръчващите системи с първа награда от 1 милион долара и над 50 хиляди участници от целия свят. След спечелването му през 2009, се провеждат редица други състезания по неговия пример - KDDCup 2011 (Dror, 2012), KDDCup 2012 (Niu, et al., 2012), ECML-PKDD 2011 Discovery Challenge (Antulov-Fantulin, et al., 2011), RecSys Challenge 2014, като прави впечатление прехвърлянето на областта на изследване от препоръчване на продукти за целите на онлайн търговия към препоръчване на информация в социалните мрежи. KDDCup2012 използва реални данни от най-голямата китайска социална мрежа Weibo¹, а RecSys Challenge2014 използва данни от социалната мрежа Twitter².

След първоначалното бурно развитие на препоръчващите системи, базирани на съдържание и сътрудничество, започват да се развиват и други видове препоръчващи системи (Adomavicius & Tuzhilin, 2005). Основните типове препоръчващи системи според Бърк (Burke R. , 2007) се разделят в следните 6 категории:

- Препоръчващи системи, базирани на съдържание (content based filtering)
- Препоръчващи системи, базирани на сътрудничество (collaborative filtering)
- Препоръчващи системи, използващи демографски фактори (demographics recommendation systems)

¹ <http://www.weibo.com/>

² <http://www.twitter.com/>

- Препоръчващи системи, използващи знания (knowledge based recommendation systems)
- Препоръчващи системи, базирани на общности (community based recommender systems)
- Хибридни препоръчващи системи.

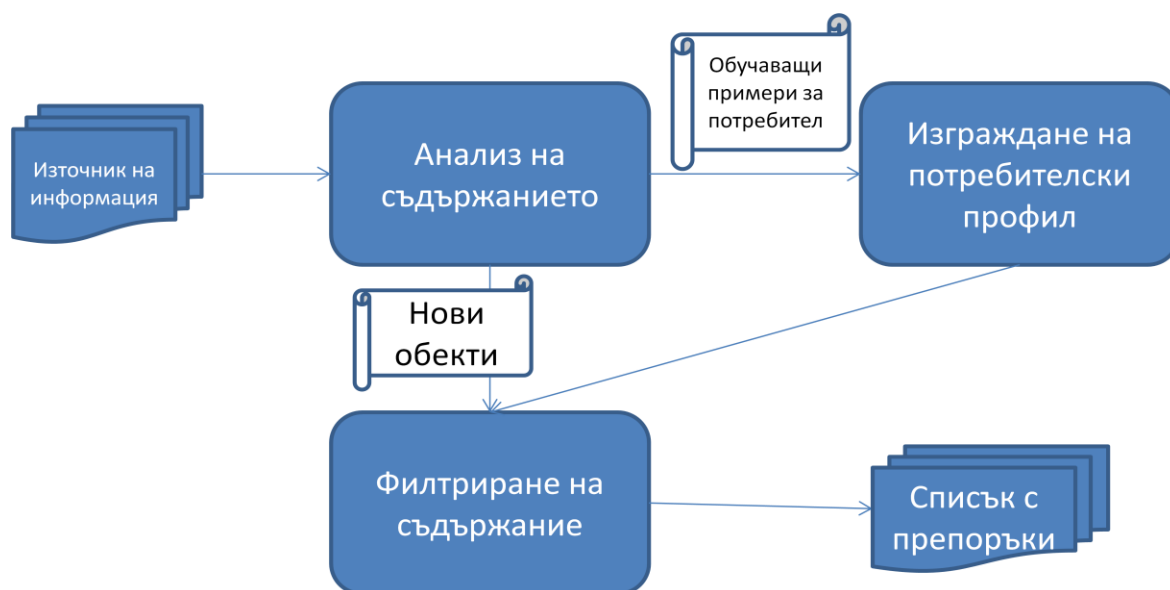
Следващите точки разглеждат различните категории препоръчващи системи, методите за оценяване, платформи и софтуерни инструменти за препоръчващи системи.

1.1. Препоръчващи системи базирани на съдържание

Препоръчващите системи базирани на съдържание се опитват да препоръчват обекти близки до вече харесаните от потребителя. Всички обекти се представят като съдържание описано от множество свойства. Харесаните от потребителя обекти изграждат потребителски профил, който се използва за филтриране на нови обекти.

Препоръчващата система, базирана на съдържание обикновено се изгражда от 3 основни модула (Pazzani M. , 1999) (Billsus & Pazzani, 2000) (Lops, de Gemmis, & Semeraro, 2011):

- Модул за анализиране на съдържанието(Content Analyzer) – този модул преобразува обектите в структурирано съдържание, което е изградено от множество от свойства и стойности за тях. Така изграденото структурирано съдържание в последствие се използва за сравнение и изчисляване на близост до потребителски профил
- Модул за изграждане на потребителски профил – потребителският профил се изгражда от структурираното съдържание на обектите харесани от потребителя.
- Модул за филтриране на съдържание – това е модулът, който изготвя препоръките. В него се изчислява близостта между структурираното съдържание на обектите и потребителския профил и в резултат се получава сортиран списък с препоръки.



Фигура 1. Архитектура на система за препоръчване на информация базирана на съдържание.

Изграждането на потребителски профил може да се разглежда като учене на класификатор (Pazzani & Billsus, 2007). Задачата е: от множество от обекти класифицирани в два класа: харесани и не, да се изгради модел, който да класифицира нов пример в един от двата класа. Възможно е да се използват различни класификационни алгоритми като представените от Пазани и Билсъс са класификационно дърво(ID3), класификация с правила(RIPPER), метод на най-близкият съсед и наивен бейсов класификатор.

Друг подход за изграждане на потребителски профил е използването на алгоритъма на Рочо за релевантна обратна връзка (Rocchio, 1971). Основната причина за успеха му е в опростеното разглеждането на документите като „торба с думи“ и изграждането на опростен потребителски профил. При този подход се изгражда потребителски профил, който е вектор от ключовите думи, съдържащи се в обектите, харесани или не харесани от потребителя.

Примери за системи използващи препоръки базирани на съдържание с вариант на алгоритъма на Рочо са описани в (Ittner, Lewis, & Ahn, 1995) (M. & Billsus, 1997) (Balabanovic & Shoham, 1997) (Ahn, Brusilovsky, Grady, He, & Syn, 2007)

В следващите секции разглеждаме формалното представяне на документите чрез векторно-пространствен модел и два подхода за изграждане на потребителски профил – с използване на наивен бейсов класификатор и с алгоритъма на Рочо.

1.1.1. Векторно-пространствен модел

За формализиране на задачата за препоръчване базирано на съдържание използваме векторно-пространствен модел (Salton, Wong, & Yang, 1975). При него всяко съдържание се представя като вектор в N-мерното пространство, където N е се определя от размера на речника с ключови думи $T = \{t_1, t_2, \dots, t_N\}$ използван за описване на съдържанието на обектите. Векторът за съдържанието на всеки обект се изписва като $d_i = \{w_{i1}, \dots, w_{iN}\}$, където w_{ik} е теглото асоциирано за ключовата дума t_k в обекта d_i . За определяне на теглата на ключовите думи обикновено се използва метриката TF-IDF, която използва, както честотата на срещане на думата в документа, така и честотата на думата в множеството от документи $D = \{d_1, d_2, \dots, d_m\}$:

$$tfidf(t, d, D) = tf(t, d) \cdot idf(d, D) \quad (1-1)$$

Функцията tf се дефинира използвайки функцията f , която връща срещанията на думата в съдържанието:

$$tf(t, d) = \frac{f(t, d)}{\max(f(t', d): t' \in d)} \quad (1-2)$$

Функцията idf се дефинира с формулата

$$idf(t, D) = \log \frac{|D|}{count(t, D)} \quad (1-3)$$

където $count(t, D)$ връща броя на документите от корпуса D, в които думата t се среща поне веднъж.

Използвайки векторно-пространствен модел близост между обектите се търси като близост между вектори. Една от най-разпространените формули за това е формулата за косинусова прилика:

$$\cos(d_a, d_b) = \frac{d_a \cdot d_b}{|d_a| |d_b|} = \frac{\sum_{i=1}^m d_{a,i} d_{b,i}}{\sqrt{\sum_{i=1}^m d_{a,i}^2} \sqrt{\sum_{i=1}^m d_{b,i}^2}} \quad (1-4)$$

1.1.2. Наивен бейсов класификатор

Препоръчването на информация може да се разгледа от гледна точка на задачата за класификация: въз основа на историческа информация какви обекти харесва потребителя, да се изгради модел, който да разделя новите обекти в два класа - такива, които ще се харесат на потребителя и такива, които няма да му се харесат. Дефинирайки така проблема може да използваме стандартен алгоритъм за класификация като наивен бейсов класификатор (McCallum & Nigam, 1998).

Наивният бейсов класификатор се базира на формулата за условна вероятност на Бейс:

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)} \quad (1-5)$$

В нашият случай се търси при условие, че имаме документ d кой клас е най-вероятен за него. За наивният бейсов класификатор се прави едно „наивно“ допускане, че всички ключови думи са условно независими една от друга, като по този начин вероятността да имаме документ d в клас c се свежда до произведение от вероятностите за ключовите думи да се срещат в конкретния клас

$$P(d|c) = \prod_{t_i \in d} P(t_i|c) \quad (1-6)$$

За намиране на класа следва да се намери, кой клас има най-голяма вероятност за конкретния документ и множеството му от термини:

$$class = \max_c (P(c) \prod_{t_i \in d} P(t_i|c)) \quad (1-7)$$

1.1.3. Подход базиран на алгоритъма на Рочио

Алгоритъмът на Рочио (Rocchio, 1971) оригинално се използва за извличане на релевантни документи, използвайки обратна връзка от потребителя за вече направени търсения (relevance feedback). Всеки документ се разглежда като „торба с думи“ и изгражда класификационен модел (потребителски профил) състоящ се от всички харесани от потребителя документи. Теглото на всяка ключова дума се определя със следната формула:

$$w_{ki} = \beta \sum_{\{d_j \in POS_i\}} \frac{w_{kj}}{|POS_i|} - \gamma \sum_{\{d_j \in NEG_i\}} \frac{w_{kj}}{|NEG_i|} \quad (1-8)$$

В горната формула POS_i е множеството от обектите, харесани от потребителя i , а множеството NEG_i , е множеството от обектите, които не са харесани от потребителя i .

1.2. Препоръчващи системи, базирани на сътрудничество

Препоръчващите системи, базирани на сътрудничество започват своето развитие преди около двадесет години. Първите статии, в които се среща термина филтриране на база сътрудничество (collaborative filtering) са на проекта GroupLens (Resnick, Iacovou, Suchak, Bergstrom, & Riedl, 1994) и в изследователския

център на Ксерокс (Goldberg, Nichols, Oki, & Terry, 1992). Този вид препоръчващи системи е един от най-използваните и изследваните в началото на този век, като актуалността му се потвърждава от направеното от Нетфликс състезание за препоръчване на филми с награда един милион долара (Zhou, Wilkinson, Schreiber, & Pan, 2008). Отличен обзор на препоръчващите системи, базирани на сътрудничество може да се намери в (Schafer, Frankowski, Herlocker, & Sen, 2007) и в (Zanker, Felfernig, & Friedrich, 2011)

Основната идея на препоръките, базирани на сътрудничество е, че близките по предпочитания потребители в миналото са близки и в настоящето. При това предположение, за да направим препоръка за даден потребител е достатъчно да намерим потребители, които са харесвали същите неща като него, и да препоръчаме продуктите, които са харесани от тях, но са все още неизвестни за конкретния потребител.

Основните типове препоръки, базирани на сътрудничество са запомнящи (memory-based) и изграждащи модел (model-based). Първият тип се нарича още препоръки базирани на най-близкия съсед.

1.2.1. Сътрудничество използващо най-близкия съсед

Препоръките, базирани на сътрудничество, използващи най-близкия съсед се разделят от своя страна на два вида: най-близкият съсед на база потребители и най-близкият съсед на база продукти. Входните данни за задачата за препоръчване базирано на сътрудничество могат да се представят като матрица потребител-продукт, която съдържа оценките, дадени от потребителите за съответните продукти (Фигура 2).

	Продукт 1	Продукт 2	Продукт 3	Продукт 4
Потребител 1	5	1	1	3
Потребител 2	1	5	1	4
Потребител 3	1	1	5	3

Фигура 2. Таблица с оценки потребител-продукт

Всички алгоритми от категорията препоръки, базирани на сътрудничество работят използвайки данните от тази таблица, но подхождат към решаването на задачата от различна страна.

1.2.1.1. Най-близкият съсед на база потребители

Това е класическият подход за препоръки, базирани на сътрудничество. При него на база на исторически данни, за това какво харесват потребителите (данни в табличен вид като Фигура 2), за конкретен потребител се търсят най-близките до неговите оценки потребители. В последствие му се препоръчва продукт, който те са харесали, а той още не е оценил.

Основните въпроси, на които трябва да се отговори са:

- Каква метрика да се използва за намиране на близост между потребителите?
- Колко най-близки съседа да се разгледат?
- Кои от продуктите харесани от най-близките съседи да се препоръчат на потребителя?

При избиране на метрика за близост между потребителите трябва да се вземе в предвид, че различните потребители използват по различен начин скалата за оценяване. При използване на скала за оценяване 0-10 за някой потребител може 10 да означава филм, който му е харесал, а 6 филм, който не му е харесал докато друг потребител може да дава оценка 7-8 за филмите, които са му харесали и 0-1 за филмите, които не са му харесали. Поради тази особеност на оценяването при избирането на метрика за близост данните трябва да се използват нормализирани в нея. Широко използвана е метриката „Корелация на Пиърсън“:

$$PC(u, v) = \frac{\sum_{i \in I_{uv}} (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in I_{uv}} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{i \in I_{uv}} (r_{v,i} - \bar{r}_v)^2}} \quad (1-9)$$

При корелацията на Пиърсън нормализацията се извършва като се изважда средната оценка която дават потребителите от оценката, която са дали за конкретен продукт. Този подход дава възможност да се открият потребители с подобни интереси дори и те да използват различен подход за оценяване.

Избиране на броя на най-близките съседи, които ще бъдат използвани за препоръчване зависи от конкретната област, в която се правят препоръките. Ако се разгледат много малко съседи, ще имаме по-малък брой от продукти, които ще можем да препоръчаме. Ако пък разгледаме прекалено много съседи, изчислителната сложност се увеличава и е възможно да се внесе шум от съседи, които са избрани за

близки до потребителя, но всъщност имат доста по-различни предпочитания от него. За областта за препоръчване на филми се препоръчва използването на между 20 и 50 съседа (Herlocker, Konstan, & Riedl, 2002)

1.2.1.2. Най-близкият съсед на база продукт

Един от основните проблеми при препоръките базирани на сътрудничество е изчислителната сложност на алгоритмите. Когато броят на потребителите е много по-голям от броя на продуктите, може да се приложи оптимизация като се търси близост между продукти в матрицата потребител-продукт, вместо между потребители. По този начин се използват филмите харесани вече от потребителя и се търсят други филми харесвани от същите потребители. Този подход много прилича на препоръчването, базирано на съдържание, като разликата е, че за изчисляването на близост между продуктите се използва само информацията за това, кои потребители са харесали продукта.

За изчисляване на близост между продукти може да се използва косинусова прилика

$$\cos(p_1, p_2) = \frac{p_1 p_2}{|p_1| |p_2|} \quad (1-10)$$

При използването на косинусова прилика може да има проблем с това, че различните потребители оценяват различните продукти по различен начин. За нормализиране на данните се прилага изваждане на средната оценка давана от потребителя по подобен начин, както във формулата за корелация на Пийрсън. С това формулата, която се ползва придобива следния вид:

$$\text{sim}(p_1, p_2) = \frac{\sum_{u \in U} (r_{u,p_1} - \bar{r}_u)(r_{u,p_2} - \bar{r}_u)}{\sqrt{\sum_{u \in U} (r_{u,p_1} - \bar{r}_u)^2} \sqrt{\sum_{u \in U} (r_{u,p_2} - \bar{r}_u)^2}} \quad (1-11)$$

Направените препоръки на базата на най-близки продукти за разлика от препоръките на база на най-близки потребители правят по-стабилни препоръки, които приличат повече на продуктите, които потребителя е харесал в миналото си. Това може да е добре в някои ситуации, но липсата на изненадващи препоръки може да доведе до правенето на прекалено еднообразни предложения.

1.2.2. Изграждане на модел с латентни свойства

Широко използван подход за препоръки базирани на сътрудничество е факторизиране на матрицата потребител-продукт (Matrix Factorization). С подхода факторизиране на матрици се цели промяна на пространствения модел с прехвърлянето

на потребителите и продуктите в пространство дефинирано от латентни фактори. За целта се използва математическият модел разлагане на собствени вектори със собствени стойности (Singular Value Decomposition -SVD). При него имаме разлагане на една матрица M на 3 други с по-малка размерност $M = U\Sigma V^T$, където матрицата Σ съдържа собствените стойности на M , а U и V са съответно лява и дясна матрица със собствени вектори.

Методите, базирани на разлагане на собствени вектори със собствени стойности са се доказали като много успешен подход за препоръки базирани на сътрудничество (Koren, Bell, & Volinsky, 2009). За факторизиране на матрицата потребител-продукт обаче не може да се ползва известният математически модел понеже той не може да работи при липсващи стойности в матрицата M , а повечето стойности в матрицата с оценките потребител-продукт не са налични и точно за тяхното намиране се нуждаем от препоръчваща система. За решението на задачата най-често се ползват подходи като стохастично градиентно спускане.

За да приложим търсене с градиентно спускане дефинираме проблема по следният начин. Искаме да обучим модел, който да предсказва оценката на потребител u за продукт i по следният начин:

$$\bar{r}_{ui} = \mu + b_i + b_u + q_i^T p_u \quad (1-12)$$

Където μ е средната оценка за всички продукти, b_i е стандартното отклонение от средната оценка за продукт i , а b_u е стандартното отклонение от средната оценка за потребител u , векторите q_i и p_u се явяват съответно собствените вектори на продукта i и на потребителя u . Оптимизационната задача е да се минимизира функцията:

$$\min_{b_*, q_*, p_*} \sum_{(u,i) \in K} (r_{ui} - \mu - b_i - b_u - q_i^T p_u)^2 + \lambda_4 (b_i^2 + b_u^2 + \|q_i\|^2 + \|p_u\|^2) \quad (1-13)$$

Където λ_4 е параметър, който се нарича параметър на регулация и отговаря за това да не пораснат прекалено много стойностите на променливите, което би довело до прекалено специфициране към обучаващите данни.

При обучението с градиентно спускане се извършва итеративно обхождане по обучаващите примери и променяне на параметрите спрямо изчисленото отклонение e_{ui} от вярната оценка.

$$\begin{aligned} b_u &\leftarrow b_u + \gamma(e_{ui} - \lambda_4 b_u) \\ b_i &\leftarrow b_i + \gamma(e_{ui} - \lambda_4 b_i) \\ q_i &\leftarrow q_i + \gamma(e_{ui} p_u - \lambda_4 q_i) \\ p_u &\leftarrow p_u + \gamma(e_{ui} q_i - \lambda_4 p_u) \end{aligned} \quad (1-14)$$

Параметърът γ определя скоростта на обучение.

В състезанието „Наградата на Нетфликс” се правят експерименти с използване на неявни оценки от потребителите при факторизиране на матрицата. Този подход е наречен от изобретателите си SVD++ (Koren, 2008), като показва по-добри резултати от стандартния подход върху базата данни от състезанието. Невявните данни, които се използват е информацията за това че потребител е дал оценка за филм. Предположението в тази ситуация, е че потребителя по-някакъв начин е свързан с този филм без значение дали го е харесал или не. Понеже от всичките хиляди филми е дал оценка точно за него.

При SVD++ формулата, с която се изчислява предсказание за оценка на филм, е следната:

$$\bar{r}_{ui} = \mu + b_i + b_u + q_i^T \left(p_u + |R(u)|^{-\frac{1}{2}} \sum_{j \in R(u)} y_j \right) \quad (1-15)$$

Където с $R(u)$ обозначаваме всички филми оценени от потребителя, а у допълнителните параметри на системата, които следва да се обучат с градиентно спускане.

Друг вариант за факторизиране на матрици е с използване на времевата информация за това кога са дадени различните оценки от потребителите. За интегриране на тази информация се изменят всички параметри от горната формула да бъдат функции зависещи от времето и съответно формулата за изчисляване на предсказанията придобива следният вид:

$$\bar{r}_{ui} = \mu + b_i(t_{ui}) + b_u(t_{ui}) + q_i^T \left(p_u(t_{ui}) + |R(u)|^{-\frac{1}{2}} \sum_{j \in R(u)} y_j \right) \quad (1-16)$$

1.2.3. Сравнение между препоръките базирани на съдържание и на сътрудничество

Използването на препоръки базирани на съдържание има няколко предимства пред препоръките базирани на сътрудничество:

- Независимост на потребителите – потребителският профил е индивидуален за всеки потребител и не се влияе от профилите на другите потребители
- Прозрачност на препоръките – препоръките, базирани на съдържание са лесни за обяснение. Всяка препоръка е направена на

основа на близост на препоръчания обект до вече харесан от потребителя друг обект.

- Препоръки за нов обект – препоръките базирани на съдържание, нямат проблем при препоръчване на нов обект, който няма история на харесванията до момента. Всеки обект се представлява от своето съдържание и няма значение дали е нов или наличен от много време.

Недостатъците на препоръчващите системи базирани на съдържание са следните:

- Прекалено специализиране – препоръчващите системи, базирани на съдържание генерират само препоръки близки до вече разгледани обекти. Те не генерират различни или изненадващи препоръки, поради което с времето, интереса на потребителя може да бъде намален.
- Препоръки за нов потребител – при регистрацията на нов потребител системата трябва да натрупа данни за него, за да му изгради профил и да може да му препоръчва релевантни продукти.
- Ограничения в обработката на съдържанието –структурирането на съдържанието не е тривиална задача и често може да настъпят проблеми. Пример за проблеми са например наличието на синоними и омоними. Ако думите от съдържанието се третират като „торба с думи”(bag of words) то не се разбира истинската семантика на думите. Разглеждайки семантиката на различните съдържания се стига до отворени проблеми от областта на обработка на естествен език и семантичен анализ.

1.3. Препоръчващи системи, използващи демографски фактори

С развитие на Web 2.0 интернет потребителите все по-уверено предоставят свои демографски данни на различните платформи за споделяне на съдържание. Системите, които използват тези данни се наричат демографски препоръчващи системи, като демографските данни, които ползват са възраст, пол, местоположение, религия и др. При демографските препоръчващи системи се прави предположението, че хора от една

демографска група имат подобни предпочитания. Предимствата са им, че нямат проблем със студено стартиране при нов потребител и, че могат да предоставят лесно обяснение на препоръките си. Като недостатък може да се отбележи прекаленото обобщаване на потребителите, като ако човек е с абсолютно различни предпочитания от общите за демографската им група той ще получава съвсем нерелевантни препоръки.

Подробно разглеждане на демографските препоръчващи системи може да се намери в (Krulwich, 1997) (Pazzani M. J., 1999)

1.4. Препоръчващи системи, използващи знания

Препоръчващите системи използващи знания могат да помогнат за ситуации, в които препоръките, базирани на съдържание и сътрудничество не могат да направят препоръка. Например препоръчващите системи, използващи знания се използват в ситуации, в които няма как да съществува дълга потребителска история – покупка на жилище, кола и др. При неналичие на потребителска история не може да се приложат препоръки базирани на сътрудничество, а при ситуации, при които имаме остаряване на продукти, например електроника, не може да се приложат препоръки базирани на съдържание.

Препоръчващите системи, използващи знания се характеризират с голяма интерактивност и възможности за задаване на сложни заявки от потребителя. Основните данни се съхраняват обикновено в структури за знания хранилища за семантични знания, като набавянето на тези знания в доста случаи може да е доста скъпо. Подробно разглеждане на препоръчващите системи използващи знания може да се намери в (Burke R. , 2000) и (Felfernig, Friedrich, Jannach, & Zanker, 2006).

1.5. Препоръчващи системи, базирани на общности

Хората са склонни да се доверят повече на препоръки от приятели отколкото на препоръки, базирани на непознати хора близки до техният профил (Sinha & Swearingen, 2001) Това се определя, както от желанието на хората да получат обяснение за препоръките, които получават, така и от желанието да споделят дейности със своите приятели.

Едни от първите изследвания, които използват социална информация за връзки между потребителите се основават на мрежи на доверие. Маса и Авесани

(Massa & Avesani, 2007) показват, че използването на информация за доверие между потребителите намалява значително проблема със студения старт (cold start) при генериране на препоръки за нови потребители и допринася за подобряване на качеството на препоръчване. Те използват за експериментите си социалната мрежа за публикуване на ревюта за продукти Epinions³, която позволява на потребителите да оценяват различни продукти и да създават мрежа от потребители, на чиито оценки имат доверие.

Друг пример за мрежа с явно указване на доверие между потребителите е FilmTrust (Golbeck & Hendler, 2006), в която потребителите могат да дават оценки за филми и да дефинират доверие помежду си.

Маса и Авесани разглеждат 2 подхода, използващи доверие. Подход, който използва явно пряко доверие и подход при който се прави разпространяване на доверието за разширяване на връзките на доверие. При подхода без разпространяване на доверието се получават възможно най-точни резултати, но покритието на продуктите е ниско. Добавяйки разпространение на доверието те успяват да получат по-добро покритие на продуктите, но точността на резултатите се намаля и спада под тази на препоръките, базирани на сътрудничество. Въпреки това чрез препоръките базирани на доверие те показват, че ефективно се преодолява проблема със студения старт при нов потребител.

Веднъж дефинирано доверието между потребителите за изчисляване на предсказание за продукт може да се използва формулата за претеглено средно:

$$p_{u,i} = \frac{\sum_{u' \in R^T} t_{a,u'} r_{u',i}}{\sum_{u' \in R^T} t_{a,u}} \quad (1-17)$$

Горната формула не взема в предвид, че различните потребители оценяват по-различен начин продуктите. Маса и Авесани използват следната модификация за справяне с този проблем:

$$p_{u,i} = \bar{r}_u + \frac{\sum_{u' \in R^T} t_{a,u'} (r_{u',i} - \bar{r}_{u'})}{\sum_{u' \in R^T} t_{a,u}} \quad (1-18)$$

Където $\bar{r}_u$ е средната оценка давана от потребител.

За разпространение на доверието може да се използва следната формула

$$t_{a,u} = \frac{\sum_{v \in WOT^+(a)} t_{a,v} t_{v,u}}{\sum_{v \in WOT^+(a)} t_{a,v}} \quad (1-19)$$

³ [Http://epinions.com](http://epinions.com)

Където $WOT^+(a)$ е мрежата на доверие на потребител a при дефиницията и се използва праг на доверието, като при доверие към потребител под този праг се приема, че той не е от мрежата на доверие.

Социалните мрежи също могат да се разглеждат като мрежи на доверие като при тях доверието се указва с добавяне на явна връзка за приятелство между потребителите. В (Chen & Fong, November 2011) авторите използват Фейсбук, като мрежа на доверието, като комбинират доверието със сътрудничество между потребителите. Те показват, начин за използване на доверие в социалната мрежа Фейсбук, като за улеснение го дефинира като симетрична величина. Чен и Фонг базират изследванията си на изследванията направени от Казиенко и Мюзиял за препоръки в онлайн социални мрежи (Kazienko & Musiał, 2006)

1.6. Хибридни препоръчващи системи

Често само един подход за препоръчване не дава търсената точност, което налага комбиниране на повече от един подход в една препоръчваща система. Такъв вид препоръчваща система наричаме хибридна препоръчваща система. Бърк първи създава таксономия на различните видове хибридни препоръчващи системи (Burke R. , 2002). Той описва 7 вида хибридни системи, като в последствие разглеждани от архитектурна гледна точка те са намалени до 3 в (Zanker, Felfernig, & Friedrich, 2011):

- Монолитна хибридна система
- Паралелна хибридна система
- Конвейерна хибридна система

1.6.1. Хибридна система с монолитен дизайн

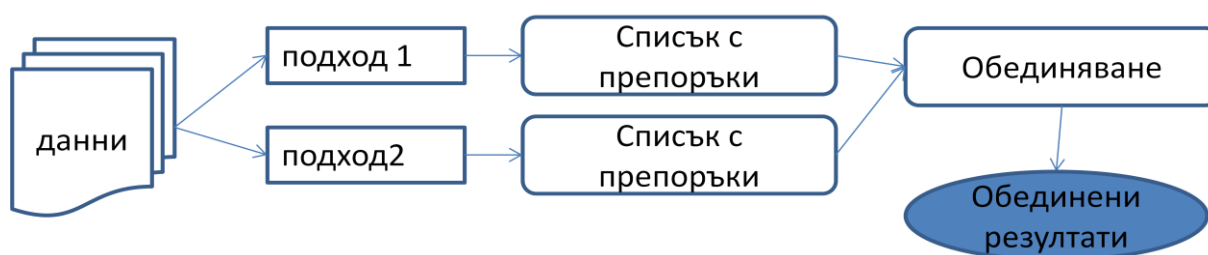
При хибридната система с монолитен дизайн имаме комбиниране на различни изгледи към данните и различни подходи чрез обединяването им в един монолитен подход. При него не се прави ясно разграничение между различните подходи, които участват в хибридната система, а по скоро имаме нов подход, използващ всички налични данни.

За прилагането на този подход следва да се направи задълбочен анализ на свойствата, които могат да се извлекат от всеки изглед на данните и използването

им от хибридният подход. Трудностите, които се забелязват са усложняване на системата и подхода за препоръчване.

1.6.2. Хибридна система с паралелен дизайн

При направата на хибридна система с паралелен дизайн имаме паралелна работа на всеки един от подходите за препоръчване, при което всеки един от тях генерира препоръки, които в последствие се обединяват с помощта на различни тегла.(Фигура 3)



Фигура 3. Архитектура на хибридна система с паралелен дизайн.

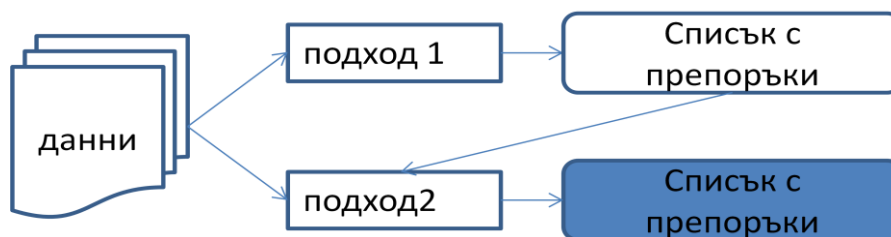
Предимството на хибридната система с паралелен дизайн е, че обединението става без да се променят различните подходи за препоръчване, което намалява сложността за имплементация. Този подход се е доказал, като най-ефективен като с него е спечелено състезанието на Нетфликс (Bell & Koren, 2007) с обединяването на над 100 различни препоръчващи подхода с тегла обучени с градиентно спускане.

1.6.3. Хибридна система с конвейерен дизайн

При хибридната система с конвейерен дизайн имаме работа на препоръчващите подходи един след друг, като на всеки следващ алгоритъм за препоръчване филтрира резултатите получени от предишния подход (Фигура 4).

Предимство на тази архитектура е, че всеки препоръчващ подход филтрира списъка с препоръки елиминирайки най-лошите и по този начин се подобрява точността на крайния списък.

При този вид архитектура всяка стъпка редуцира списъка с препоръки и не може да добавя нови елементи към него, поради което в списъка няма да се включат интересни обекти, намирани само от един от подходите.



Фигура 4. Архитектура на хибридна система с конвейерен дизайн.

1.7. Оценяване на препоръчващи системи

За да направим избор коя препоръчваща система е най-добра за конкретна област, трябва да можем формално да сравним различни препоръчващи системи и да оценим тяхната точност. Подробен анализ на различни методи за оценяване на препоръчващите системи може да се намери в (Herlocker J. , Konstan, Terveen, & Riedl, 2004) и (Shani & Gunawardana, 2011)

1.7.1. Средна абсолютна грешка

Стандартен начин за оценяване на точност на една система е да изчислим средна грешка на дадените оценки върху предварително известно тестово множество от оценки. В този случай от събраните данни за оценките дадени от потребителите образуваме две непресичащи се множества – обучаващо и тестващо. Изграждаме модел върху обучаващото множество и изчисляваме средната грешка върху тестващото множество. Формулата за средната грешка е:

$$MAE = \frac{\sum ratings |P - R|}{\#ratings} \quad (1-20)$$

В горната формула P е направеното от препоръчващата система предсказание за оценката която ще даде потребителя, а R е реалната оценка дадена от потребителя. Използваме модул при изваждането на двете оценки за да акумулираме грешката, а да не стане уравнивяване когато веднъж системата е дала оценка с X по-голяма, а след това с X по-малка.

1.7.2. Средна Квадратична Грешка

Средната Квадратична Грешка (Root Means Squared Error) е друга често използвана метрика за оценяване на препоръчващите системи. При нея грешката се акумулира след повдигане на квадрат:

$$RMSE = \sqrt{\frac{\sum ratings (P - R)^2}{\#ratings}} \quad (1-21)$$

При квадратичното акумулиране на грешката имаме по голямо наказание за по-големи отклонения от търсения резултат. Например при изчисляване MAE няма разлика в грешката ако един пример има отклонение 2 или два са с отклонение 1, докато при RMSE първият случай ще получи по-лоша оценка за грешката.

1.7.3. Метрики за точност от областта извличане на информация

За оценяване на препоръчващи системи често се използват метрики от областта на извличане на информация – точност(precision), откриване(recall). При метриката „откриване” се измерва каква част от релевантните резултати са намерени от системата:

$$R = \frac{\text{брой на релевантните документи в резултатите}}{\text{брой на всички релевантни документи}} \quad (1-22)$$

При метриката „точност” имаме измерване каква част от резултатите върнати от системата са релевантни:

$$P = \frac{\text{брой на релевантните документи в резултатите}}{\text{броят на всички документи в резултатите}} \quad (1-23)$$

Двете метрики се противопоставят една на друга като при увеличаване на броя на извлечените резултати обикновено се увеличава откриването, но се намалява точността.

F-мярка е метрика, която комбинира точност и откриване в една оценка. Това ни позволява по-лесно сравнение на различни подходи върху едни и същи данни. Най-използваната формула е за F_1 мярка:

$$F_1 = \frac{2PR}{P + R} \quad (1-24)$$

Един от недостатъците на метриките точност и откриване е, че те се измерват върху всички резултати. Когато искаме да измерим точност само върху първите n резултата се прилага формулата:

$$P@n = \frac{N_{r@n}}{n} \quad (1-25)$$

$N_{r@n}$ е броят на релевантните резултати в първите n резултати.

1.7.4. Средно Аритметична точност

Друга много популярна метрика в извличането на информация е средно аритметична точност (Mean Average Precision). При нея имаме оценяване на комбинирана точност от множество заявки. Формулата за изчисляването и е:

$$MAP = \frac{\sum_{q=1}^Q AveP(q)}{Q} \quad (1-26)$$

Където Q е броя на различните заявки, $AveP(q)$ е средната точност за заявка номер q , а средна точност изчисляваме с формулата

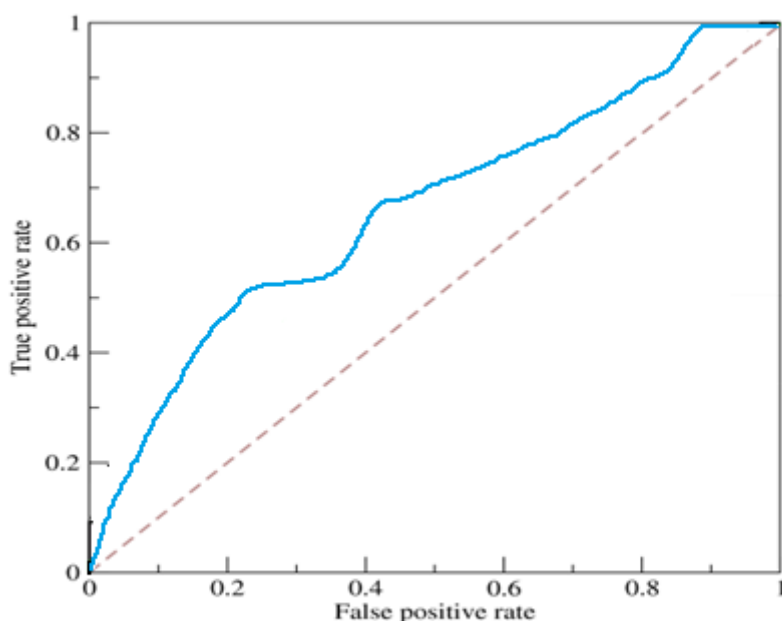
$$AveP = \frac{\sum_{k=1}^n P@k * rel(k)}{\#relevant docs} \quad (1-27)$$

$rel(k)$ е 1 ако документа на позиция k е релевантен и 0 в противен случай. Важно е да се отбележи, че за разлика от разгледаните до сега метрики, за метриката средна точност има значение подредбата на списъка от документи, като колкото по напред са релевантните резултати толкова по-голяма е средната точност.

1.7.5. РОК-крива

Друга много използван подход за оценяване е графичният подход с рисуването на РОК-крива (ROC curve). При този подход изграждаме графика в която показваме как се изменя съотношението между вярно отбелязаните за релевантни резултати към грешно отбелязаните за релевантни. Подобен вид графика е удобен за калибриране на параметрите на алгоритъм. Като например ако от параметъра зависи колко резултата ще бъдат върнати, то с промяна на параметъра в посока увеличаване на броя на резултатите се започва да се отбелязват на графиката каква част от релевантни резултати са намерени и каква част от нерелевантните (Фигура 5).

При сравнение на два алгоритъма с използването на РОК-крива обикновено се сравнява лицето на областта под кривата, като колкото е по-голямо толкова по-добър е алгоритъмът.



Фигура 5. Примерна РОК-крива

1.7.6. Среден Реципрочен Ранк

Средният реципрочен ранк (Mean Reciprocal Rank) е метрика за оценяване, която се използва за осреднена оценка на множество от заявки за това кой под ред е първият релевантен резултат измежду върнатите резултати.

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \quad (1-28)$$

Оценката среден реципрочен ранк дава информация за това доколко системата връща максимално нагоре в списъка релевантен резултат, но не взема в предвид колко са релевантните резултати върнати от системата.

1.7.7. Нормализирана обезценена кумулативна печалба

Обезценената кумулативна печалба (Discounted Cumulative Gain) е метрика, която оценява качеството на подредбата на списък с резултати. За изчисляването и се използват следните две помощни функции

- $disc(r(i))$ – функция, която връща коефициент на обезценяване за позицията $r(i)$ на резултата i .
- $u(i)$ – функция, която връща оценка за релевантността на резултата i .

Използвайки горните две функции формулата за обезценена кумулативна печалба е следната:

$$DCG(r) = \sum disc(r(i))u(i) \quad (1-29)$$

Горната формула не е нормализирана. За да се нормализира се използва оценката която се получава ако се сметне обезценена кумулативна печалба върху перфектно наредена поредица от резултати. Формулата за нормализирана обезценена кумулативна печалба е следната:

$$nDCG(r) = \frac{DCG(r)}{DCG(r_{perfect})} \quad (1-30)$$

1.7.8. Онлайн оценяване

Оценяването върху предварително събрани данни не винаги дава коректна информация доколко дадена препоръчваща система ще се хареса на потребителите. Това е така понеже събраните данни са повлияни от това какво потребителят е видял и какво е оценил. Възможно е в историческите данни да има още много обекти, които биха били интересни за потребителя, но за тях да нямаме оценка. Липсата на оценка за тях в тестовото множество, ще доведе до намаляване на оценката на препоръчващата система, понеже според събраните данни потребителя не е отбелязал, че ги харесва. Поради тази причина често след като се направи офлайн оценяване и се избере добър алгоритъм следва да се направи и онлайн оценяване за да оцени истинската точност на алгоритъма.

При онлайн оценяването се използват вече изредените метрики, а също така може да се изследват и допълнителни свойства на системата, разгледани като част от тази точка. Оценките от потребителите може да се събират неявно, чрез следене на активността им спрямо направените препоръки или може да се изиска явно, чрез попълване на въпросник. Събирането на явна информация е много „по-скъпо“ от неявната информация, но резултатите от него са по-точни и могат да допринесат за по-точно разбиране на потребителите и съществени подобрение на системата.

1.7.9. Свойства на препоръките

Когато се прави оценяване на препоръчваща система освен разгледаните до тук метрики има и още свойства, които са важни за реално внедряване на препоръчващата система. Оценяването на различни свойства на системата е важно за разбиране на нейната полезност и възможностите за нейната употреба и

развитие. Подробен обзор на различните свойства на препоръчващите системи може да бъде намерен в (Shani & Gunawardana, 2011). В следващите точки разглеждаме тези от свойствата, които са избрани като най-важни за дисертацията.

1.7.9.1. Покритие

Покритието (Coverage) е свойство, което показва за каква част от продуктите или потребителите имаме препоръки, т.е. каква част от наличните продукти се препоръчват на потребителите. Голям процент на покрити продукти означава, че системата препоръчва на потребителите дори продукти, които не са чак толкова известни и по-този начин успява да ги популяризира. Процентът на покрити потребители показва за каква част от потребителите системата може да направи препоръки. И двата вида покритие следва да се комбинират с измерване на точност на препоръките, за да се види каква част от препоръките дадени от системата са релевантни.

1.7.9.2. Доверие

Доверието на потребителите е важно свойство, от което зависи доколко успешна ще бъде препоръчващата система. Ако потребителите нямат доверие в препоръките на системата, то те не биха ги използвали и биха предпочели да не им се препоръчват продукти за да не им се губи времето.

Един от начините за спечелване на доверието на потребителите е в началото да им се препоръчват само продукти, които са много вероятни да им харесат – популярни продукти и много подобни на продуктите, които потребителя вече е харесал. Изграждайки с времето достатъчно доверие в системата, те вече може да почнат да получават препоръки и за по-непопулярни продукти.

Предоставяне на обяснение за препоръките, направени от системата също допринася за увеличаване на доверието в нея. Например, когато потребител разбере, че даден продукт му е препоръчан понеже X на брой потребители са го закупили след друг закупен продукт, той би го разгледал понеже ще разбере, че може да му е от полза.

Доверие в системата може да се изгради и чрез предоставяне на интерфейс за потребителя да настройва своята препоръчваща система. По-този начин той ще

се чувства удовлетворен за това, че има възможността сам да определя от какво се интересува.

Доверието в системата, често се измерва с анкетиране на потребителите (Herlocker, Konstan, & Riedl, 2000) (Pu & Chen, 2006).

1.7.9.3. Новост

Свойството новост(novelty) показва доколко системата препоръчва на потребителя продукти, които не са му известни. Пример за такъв вид препоръка е препоръката на нов филм, в който играе любимият актьор на потребителя, но потребителя още не знае за него. Често потребителят знае за някои продукти от източници различни от препоръчващия сайт за това системата има по-голям шанс да предостави препоръка-новост за по-непопулярни продукти.

1.7.9.4. Изненадващи препоръки

Възможността на препоръчващата система да генерира изненадващи препоръки(serendipity) е нейно важно свойство, което използвано внимателно може да привлече потребителя да използва по-активно системата. Изненадващи препоръки наричаме препоръки на продукти, които се харесват на потребителя въпреки, че до момента той не е показвал на системата, че харесва подобен продукт. Основният метод за генериране на подобен вид препоръки са препоръките, използващи сътрудничество, като трябва да се внимава как точно ще бъде представена препоръката. В зависимост от представянето на препоръката потребителя може да бъде увлечен в използване на препоръчващата система или може да бъде отблъснат от съображения за навлизане в личното му пространство или използване на личните му данни.

1.7.9.5. Разнообразност

Препоръчващата система трябва да предоставя разнообразни продукти на потребителите. Ако системата предоставя списък от продукти, които са всичките много подобни един на друг, то потребителя бързо може да се отегчи и да спре да я използва. Разнообразието може да бъде измерено чрез сравнение на близостта между продуктите предлагани от системата. Колкото по-голямо разнообразие има сред направените препоръки, толкова по-голям е шансът да се препоръча на потребителя нещо ново или изненадващо.

1.7.9.6. Полезност

Свойството полезност на препоръките показва как са се увеличили продажбите или печалбите на сайта заради използването на препоръчващата система. Това свойство е трудно за измерване поради сложността да се определи дали точно препоръчващата система е повлияла на потребител да закупи даден продукт или той е щял да го закупи и без нея. Въпреки трудността за измерването му то е едно от най-важните свойства за препоръчващите системи, ориентирани към електронна търговия.

1.7.9.7. Устойчивост

Свойството устойчивост измерва доколко системата се влияе от атаки срещу нея, целящи да намалят или увеличат оценката на даден продукт. Пример за такава атака е използването на автоматични програми за даване на оценки за продуктите целящи да повишат оценката и полярността на даден продукт. Такъв вид атаки са широко използвани за измамно увеличаване на рейтинга на продукти от недобросъвестни разпространители.

1.7.9.8. Адаптивност

Адаптивността е свойство, което показва доколко системата може да предостави адекватни препоръки при временна или постоянна промяна в интересите на потребителя. Например възможно е обект да стане много популярен, и за момент да стане интересен и за потребител, който никога не се е интересувал от подобен вид продукт. В последствие обаче потребителя може да се върне към предишните си предпочитания и да не се интересува повече от подобен вид продукти. Възможно е да се получи и другата крайност, с която този вид продукти остава интересен за потребителя и той да промени предпочитанията му. В (Koychev & Schwab, 2000) Койчев показва различни подходи за адаптиране към променящите се потребителски интереси. В (Koychev, 2000) Койчев показва детайли относно използване на метод за забравяне за борба с промяна на потребителските предпочитания и смяна на концепциите.

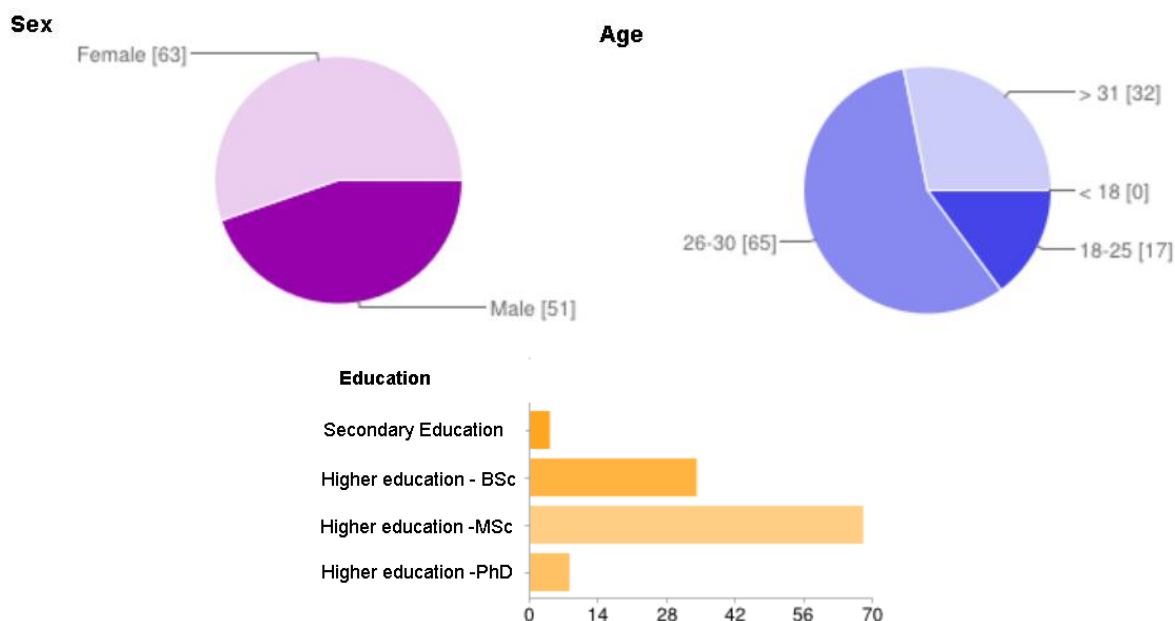
1.7.9.9. Разширяемост

Свойството разширяемост показва доколко препоръчващата система може да се справи с увеличаване на броя на потребителите и продуктите. Това свойство често се оценява с изчислителната сложност на алгоритмите използвани от системата. Свойството разширяемост е важно понеже основната задача на препоръчващите системи е да помагат на потребителите при постоянно увеличаващ се брой продукти. При подобряване на системата се очаква, че тя ще стане и по атрактивна и ще се увеличат и броя на потребителите.

Глава 2. Изследване и анализ на проблема

2.1. Изследване на необходимостта на потребителите

За целите на дисертационният труд е направен подробен анализ на нуждите на потребителите в най-популярната социална мрежа Фейсбук. Проблема за препоръчване на съдържание на потребителите на социалните мрежи е изследван от гледна точка на необходимостта на потребителите от система за препоръки и от гледна точка на това каква точно функционалност би била в полза за потребителите. Изследването е направено чрез анкета, в която взеха участие 114 участници от 5 държави. Участниците в анкетата са със следните демографски характеристики: 70% са между 18-30 годишна възраст, а 30% са над 30 годишни. 3% са със средно образование, 30% са с образователна степен бакалавър, 60% с образователна степен магистър, 7% с образователна степен доктор. 45% от анкетираните са мъже, 55%-жени.



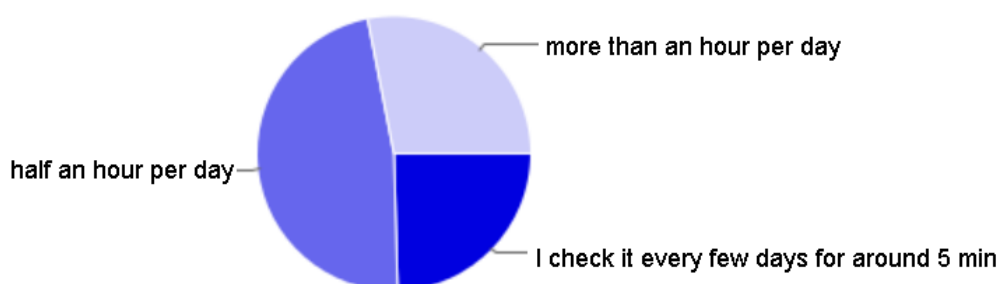
Фигура 6. Демографска информация за участниците в анкетата

Анкетата е направена, за да уточни необходимостта от ново приложение за препоръчване на информация и разглеждане на новини.

Попитани за времето, което прекарват във Фейсбук, 25% отговарят, че не са много редовни потребители и влизат в социалната мрежа от време на време за по пет минути, но останалите 75% от потребителите отговарят, че го ползват

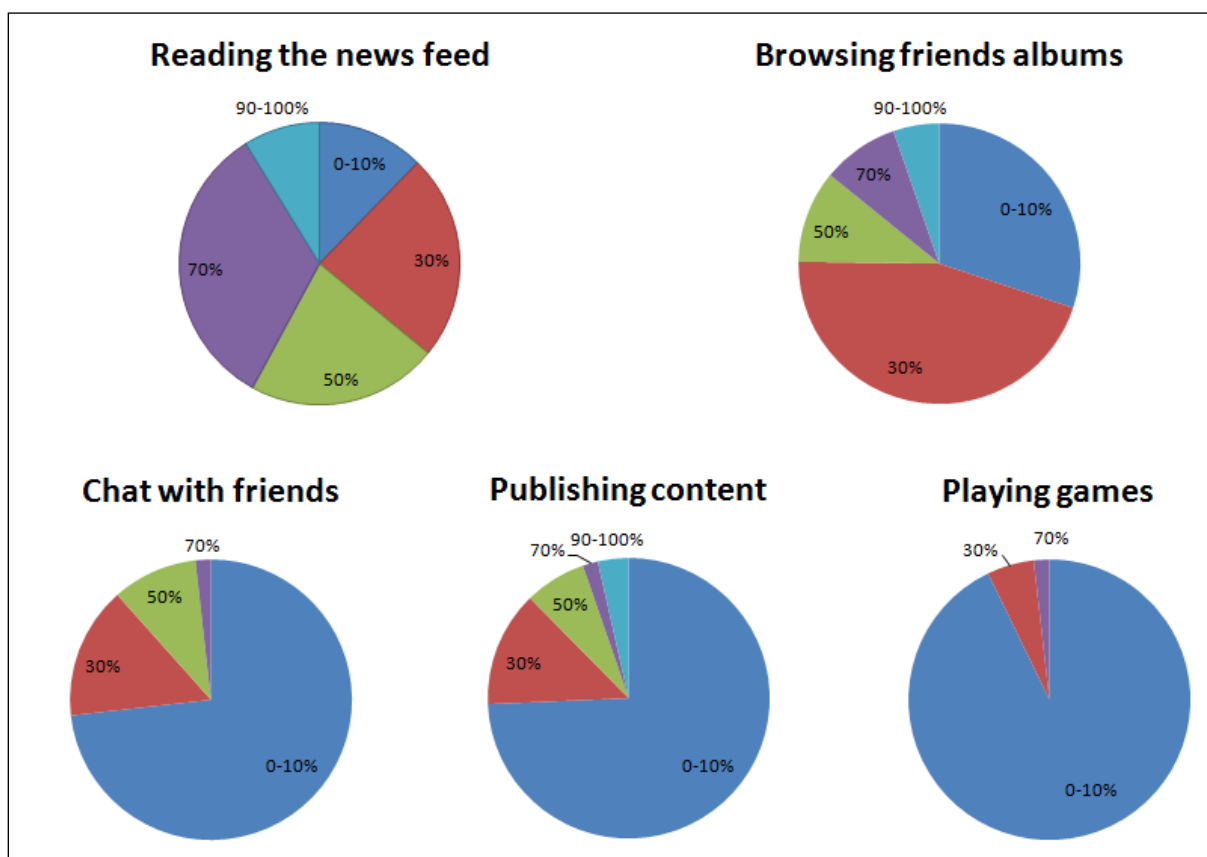
ежедневно, като 47% отделят около половин час на ден, а 25% по-повече от час на ден.

How much time do you spend at Facebook?



Фигура 7. Време прекарвано във Фейсбук от участниците в анкетата.

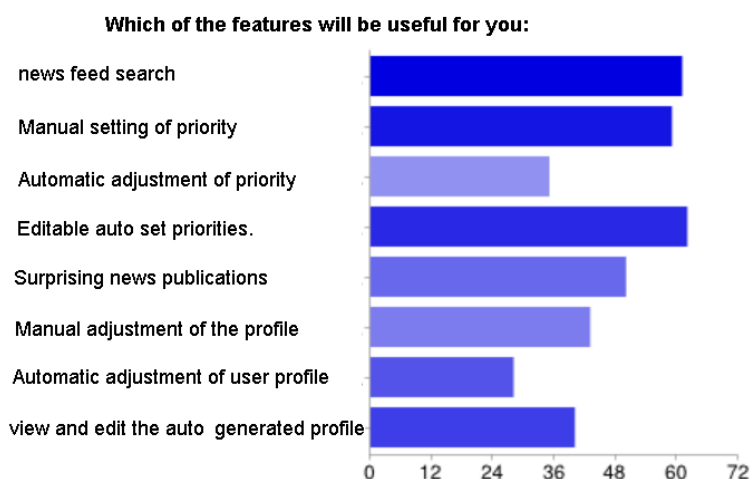
Следващата серия от въпроси, зададени на потребителя, цели да уточни, за какво потребителите използват социалната мрежа. Въпроса: „Какъв процент от времето си прекарано в социалната мрежа отделяте за конкретна активност?“ е зададен относно различните дейности в социалната мрежа: четене на потока с новини, разглеждане на снимки, публикуване на съдържание, комуникация с приятели, занимаване с онлайн игри.



Фигура 8. Време, отделяно от участниците в анкетата, за различните дейности във Фейсбук.

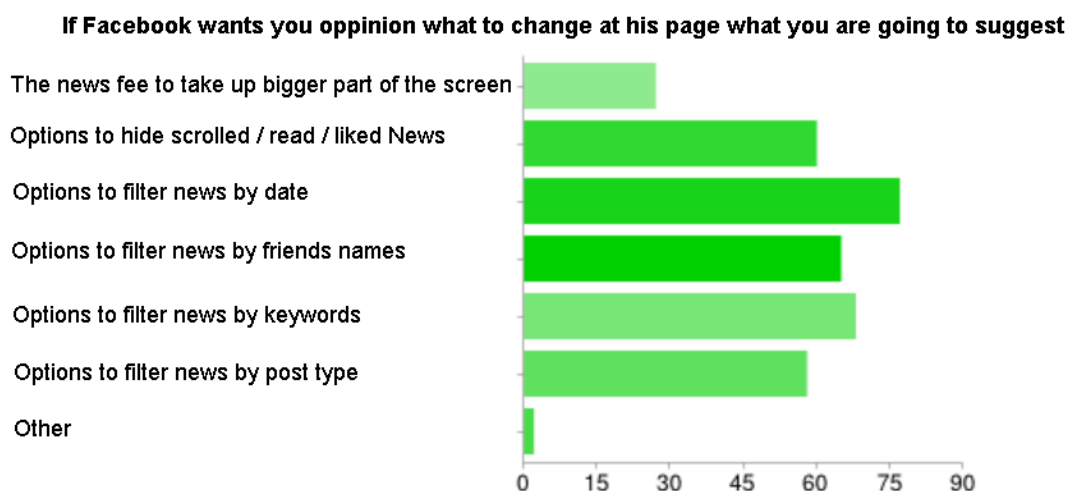
Потребителите определят „четенето на потока от новини“ като активността, която заема най-голям процент от времето им прекарано във Фейсбук, като 62% от хората определят, че тя заема повече от половината от времето, което прекарват във Фейсбук(детайлни резултати може да видите на фиг.8). На второ място по отделено време във Фейсбук се нарежда разглеждането на снимки, като 70% от потребителите твърдят, че отделят повече от 30% от времето си за тази дейност. Според тези цифри мнозинството от потребителите отделя по малко от 20% за всички останали дейности във Фейсбук.

Запитани относно функционалността, която биха искали от едно ново приложение за разглеждане на потока с новини, 54% от хората отговарят, че имат необходимост от улеснено търсене в новините(по-подобие на Гугъл). 52% отбелязват като желана възможността да се указва ръчно приоритет на новините от различните източници на информация, като само 35% искат да имат и автоматично подреждане на източниците от системата. При наличност на автоматично подреждане на източниците от системата 54% искат да имат възможност да променят автоматично зададените тегла. Само 25% от запитаните считат за подходящо системата автоматично да променя теглата в последствие, базирайки се на тяхната активност с другите потребители.



Фигура 9. Избор на нова функционалност от Фейсбук потребителите.

Допълнително потребителите показват високо желание за различни видове филтриране на новините – филтриране на прочетени/непрочетени, харесани, филтриране по дата, по приятел, по тип на съобщението и др.



Фигура 10. Избор на опции за потока с новини във Фейсбук.

Потребителите показват доста слаба активност за добавяне на нови приложения в социалната мрежа – 87% твърдят, че не са добавяли ново приложение в последния месец, но въпреки това 69% показват готовност да добавят ново приложение, ако то им даде възможност за лесно търсене и филтриране на новини, и автоматични и ръчно настройване на приоритета на новините, идващи от различните източници на информация.

От получените отговори ясно се отбелязва, че има необходимост от приложение, подпомагащо потребителя с потока на информация в социалната мрежа (69% от хората биха ползвали такова приложение). Въпреки големият процент от участниците в анкетата заявили желание да ползват подобно приложение прави впечатление, че само 25% биха се доверили приоритетите им да бъдат калибрирани автоматично и голяма част (54%) искат да могат ръчно да променят приоритетите на източниците, което показва, че потребителите са като цяло са с недоверие към препоръчващите системи и искат да имат по-голям контрол върху предоставяните им препоръки. Въпреки скептицизма на потребителите към автоматична препоръчваща система, необходимостта от по-лесно филтриране и търсене на информация в социалната мрежа определя възможността за лесно разпространение на подобно приложение и запознаване на потребителя с различните подходи за препоръчване на информация.

2.2. Свързани изследвания

Препоръките в социалната мрежа използват данни за връзките между потребителите. Тези връзки в много изследвания са моделирани като връзки на доверието. Изследванията в областта на препоръчващите системи базирани на доверие определят два вида доверие между потребителите – явно и неявно. Пример за мрежа на доверието, в която имаме явно указано доверие е мрежата за ревюта Epinions. Маса и Авесани (Massa & Avesani, 2007) изследват разпространението на доверието и недоверието в мрежата. Изследването им показва, че подходите базирани на доверие имат по-малка зависимост от обема на данните и работят, дори и когато имаме сравнително малко историческа информация за потребителя.

Социалната мрежа Фейсбук също може да се разглежда като мрежа на доверието, при нея обаче, за разлика от Epinions, нямаме явно задаване на доверие. За извличане на доверието на потребител към негов приятел следва да се използва показаната активност към публикуваното съдържание от него. В (Chen & Fong, 2011) авторите използват Фейсбук, като мрежа на доверието, като комбинират доверието със сътрудничество между потребителите. Други препоръчващи системи за Фейсбук са показани в (Bostandjiev, O'Donovan, & Höllerer, 2012) (Noel, et al., 2012) (Quijano-Sanchez, Recio-Garcia, & Diaz-Agudo, 2011), като в (Bostandjiev, O'Donovan, & Höllerer, 2012) авторите прилагат хибриден подход и акцентират върху предоставянето на обяснения за препоръките и визуализацията им. В (Noel, et al., 2012) авторите описват приложение във Фейсбук, използващо за препоръчване социално факторизиране на матрици, при което се изгражда модел подобен на препоръките със сътрудничество, но използващ и социалните елементи на мрежата. Санчес и Гарсия (Quijano-Sanchez, Recio-Garcia, & Diaz-Agudo, 2011) представят препоръчваща система, която препоръчва филми на групи от потребители. Сервия и Редондо (Servia-Rodríguez, Díaz-Redondo, Fernández-Vilas, & Pazos-Arias, 2013) представят интересен подход за определяне на силата на връзките във Фейсбук, използвайки широк набор от потребителска активност и предсказвайки в следствие от какви услуги се нуждае потребителя.

2.3. Изводи

В обобщение на направеното изследване могат да се направят 3 извода:

- Има необходимост от приложение подпомагащо потребителите в избора им на новини от социалната мрежа.
- потребителите са с недоверие към препоръчващите системи за това е необходимо приложението да обяснява резултатите си и да съдържа опции за ръчна настройка.
- За по-голяма атрактивност, приложението е добре да предлага и опции за търсене на информация

Глава 3. Препоръчване в социалната мрежа

В тази глава разглеждаме конкретните подходи и техните подобрения изследвани, като част от дисертационният труд. Основният акцент е даден на препоръките базирани на доверие, препоръки базирани на съдържание, препоръки използващи сътрудничество и хибридни модели.

3.1. Препоръки базирани на съдържание

За препоръчване на информация във Фейсбук са направени експерименти с препоръки базирани на съдържание. Препоръките базирани на съдържание работят с векторно-пространствен модел, като обекти се разглеждат като вектори от тяхното съдържание, а за потребителя се изгражда потребителски профил базиран на съдържанието на обектите с които е харесал, коментирал или публикувал. При изграждането на векторите за обектите и потребителският профил се използват текстовото съдържание описващо обектите. За оценки на ключовите думи, от които се изграждат векторите със съдържание, се използва метриката tf-idf, като допълнително се използват речници за премахване на стоп думи.

За изграждането на профил на потребителя се използва алгоритъма на Рочю, като допълнително се поставя ограничение за извличане на максимален брой ключови думи от документ. Стойността на този параметър е зададена да е $K=10$, като тази стойност е избрана експериментално с изследване на промяната в точността при различни стойности за K . Ключови думи от всеки обект свързан с потребителя се акумулират в един вектор – профил на потребителя. За оценяване на нов обект се сравнява вектора със съдържание му и създаденият потребителски профил, като за метрика за сравнение на двата вектора използваме формулата за косинусова прилика:

$$\text{similarity}(d_a, u) = \frac{d_a \cdot u}{|d_a| |u|} = \frac{\sum_{i=1}^m d_{a,i} u_i}{\sqrt{\sum_{i=1}^m d_{a,i}^2} \sqrt{\sum_{i=1}^m u_i^2}} \quad (3-1)$$

Във формулата u е вектора на профила на потребителя, d_a е вектора на документ, m е размерността на векторното пространство, $d_{a,i}$ е tf-idf оценката за

ключова дума i в документа a и u_i е tf-idf оценката за ключовата дума i в профила на потребителя u .

3.2. Препоръки базирани на сътрудничество

При изготвяне на препоръки базирани на сътрудничество за социална мрежа от затворен тип като Фейсбук, следва да се вземат в предвид следните ограничения:

- Поради защитата на личните данни в социалната мрежа, приложението има достъп само до данните на регистрираните потребители и техните приятели.
- Не бива да препоръчваме на потребител новина, до която няма достъп. Възможно е приложението да изтегли новина използвайки достъп от потребител X и тази новина да е много интересна за потребител Y , но тази новина не е предназначена за него и приложението не може да му я покаже.

Поради ограниченията написани по-горе следва, че при изготвяне на препоръки за потребител трябва да използваме само данните събрани от приложението, с правата дадени му от потребителя. Това са публикуваните от него и неговите приятели информационни обекти. Тези обекти са видими от ограничен брой потребители – потребителя, неговите приятели и приятелите на неговите приятели, като това са потребителите, които ще използваме за извършването на препоръки базирани на сътрудничество.

От изброените видове препоръки базирани на сътрудничество в глава 1 ще използваме класическият вариант за търсене на най-близък съсед на базата на потребител, като по този начин за всеки потребител на приложението списъкът от потребители за търсене на най-близки съседи е различен.

Данните, които използваме за направата на препоръки базирани на сътрудничество са всички обекти, до които има достъп потребителя и потребителите, които са ги харесали или коментирали. Тези данни могат да се представят в табличен вид (Фигура 11). В Таблицата ако обект е харесан или коментиран се отбелязва с 1, в противен случай с 0.

	item1	item2	...	itemN
user1	0	1	...	0
user2	1	0		1
....	...			
userM	0	0	...	1

Фигура 11. Таблица на харесването на социалните обекти от потребителите.

Размерността на таблицата на харесванията на обекти от потребители M се определя от броя на потребителите и обектите, които участват в нея. Потребителите са подмножество на приятелите и приятелите на приятелите на потребител. Средно един потребител има около 300 приятели, като допълнително всеки един от тях се очаква да има също 300 приятели. По този начин броят на очакваните потребители е около $300 \cdot 300 = 90,000$ потребители. Обектите, които ще се препоръчват на потребителя са измежду обектите на приятелите му, до които има достъп. От социалната мрежа изтегляме по 1000 обекта от приятел, които са харесани поне веднъж. Като по този начин включваме в таблицата около $300 \cdot 1000 = 300,000$ обекта.

От описанието на входните данни се вижда, че броят на потребителите е по нисък от броят на обектите, което мотивира избора на сравнение на потребители пред сравнение на обекти.

Всеки потребител се представя като вектор от обектите, които е публикувал харесал или коментирал, като близост между потребителите изчисляваме спрямо формулата:

$$\text{similarity}(\text{user}_k, \text{user}_l) = \frac{\sum_{i=1}^N M_{ki} M_{li}}{\sqrt{\sum_{i=1}^N M_{ki}} \sqrt{\sum_{i=1}^N M_{li}}} \quad (3-2)$$

В горната формула M_{ki} е оценката дадена от потребител k за обект i .

Използвайки тази формула за близост най-близки до потребител X ще бъдат потребители, които са харесали най-много обекти заедно с него и същевременно нямат голямо количество други харесвания.

Следва да е отбележи, че при прилагането на препоръки базирани на сътрудничество в социалната мрежа Фейсбук, не използваме често използваната корелация на Пърсън, понеже оценките давани във Фейсбук са унарни и не следва да бъдат нормализирани. Единствената нормализация, от която се нуждаем, е да се ограничи близостта до „спам“ потребители. Това са потребители, които харесват произволно хиляди неща, като дори и някаква част от тези неща да са харесвани и от потребител X , то те не бива да са близки до него. Формулата, която използваме прави тази нормализация разделяйки на броя на харесаните от потребител обекти.

За препоръчването на обекти се взимат K най-близки съседи, където K е параметър, който трябва да се специфицира в системата. При направата на различни експерименти с различни K , беше забелязано, че за препоръки на различните потребители се постига най-висока точност при различно K . Тези наблюдения обвързаха избора на K с броят на потребителите на потребител, като K е избран да е равен на половината от броят на приятелите на потребителя. Трябва да се отбележи, че най-близките потребители се избират не само от приятелите на потребителя, но и от приятелите на неговите приятели. По този начин голяма част от пасивните и не чак толкова интересните за потребителя приятели се заместват от „автоматични“ приятели, които не са му преки приятели, но са приятели на негови приятели и харесват подобно съдържание.

Веднъж генерирано множеството от най-близки съседи се съхранява в системата, като за всеки съсед се запазва неговият идентификационен номер и близостта му до потребителя. Преизчисляване на най-близките съседи се прави всяка нощ, когато се обновяват данните за всички потребителите на системата.

За препоръчване на нов обект на потребителя се оценяват обектите харесани от неговите съседи. Оценката за обектите се сформира от оценката на потребителите харесали и коментирали обекта.

$$similarity(u, d) = \frac{\sum_{u' \in neighbourhood(u) \& u' \in d} similarity(u, u')}{K} \quad (3-3)$$

3.3. Препоръки базирани на доверие

В мрежите базирани на доверие като Epinions и FimTrust средният брой връзки на доверие за потребител е 1.7 (Massa & Avesani, 2007). Това предполага, че потребителите са отбелязали доверието си само за хора, на които безрезервно се доверяват. За разлика от тях в социалната мрежа Фейсбук имаме връзки на приятелство с много по-голяма гъстота. Средно потребителите в мрежата имат около 300 приятели, което не позволява тези връзки да се използват пряко като връзки на доверие и налага връзките на приятелството да бъдат оценени и сортирани спрямо степента на доверие на потребителя в приятелите му.

За да се изчисли степента на доверие се използва информация за активността на потребителя и взаимодействието с приятелите му. Чен и Фонг (Chen & Fong, November 2011) дефинират доверието като симетрична променлива, но в реалния свят доверието често е асиметрично за това в дисертацията го дефинираме като асиметрично величина. За дефинирането на доверие на потребител в негов приятел използваме знанията за това, какво е публикувал неговият приятел, каква част от публикациите са харесани, коментирани и споделени от потребителя и каква е съвместната им активност върху други публикации. Формулата, с която правим изчислението за доверие е следната:

$$trust_{ij} = \frac{\alpha \cdot like_{ij} + \beta \cdot comment_{ij} + \gamma \cdot share_{ij}}{published_j} + \mu \cdot \frac{activity_{ij}}{activity_i} \quad (3-4)$$

Във формулата с $trust_{ij}$ отбелязваме доверието на потребител i в потребител j , $like_{ij}$ е броя на обектите публикувани от j и харесани от i , $comment_{ij}$ е броят на обектите публикувани от j и коментирани от i , $share_{ij}$ са обектите публикувани от j и споделени от i , $published_j$ е броят на обектите публикувани от j , $activity_i$ е броят на обектите, с които потребителят i е имал взаимодействие, $activity_{ij}$ е броят на обектите харесани едновременно от потребителите i и j . Параметрите във формулата $\alpha, \beta, \gamma, \mu$ определят силата на различните активности за определяне на доверието към друг потребител. Тези стойности са индивидуални за всеки потребител, като техният избор се прави автоматично с използване на процедура за градиентно спускане. Модела се обучава така, че параметрите $\alpha, \beta, \gamma, \mu$ да имат стойности, които максимизират доверието в харесаните от потребителя обекти.

За изчисляване на доверие на потребител i в обект x се използват оценките за доверие на потребителя в автора на публикацията и потребителите, които са я харесали или коментирали. Изчислението се извършва със следната формула:

$$trust_{ix} = \frac{trust_{i,author_x} + \sum_{j \in \{users \text{ interacted with } x\}} trust_{ij}}{\max(trust_i)} \quad (3-5)$$

В горната формула с $author_i$ е обозначен потребителя публикувал обект x , сумата е по всички потребители асоциирани с обекта x , а с цел нормализация разделяме на $\max(trust_i)$, което е максималното доверие което може да получи един обект. Максималното доверие е равно на акумулираното доверие от всички приятели на потребителя:

$$\max(trust_i) = \sum_{j \in friends_i} trust_{ij} \quad (3-6)$$

3.4. Отчитане на времеви фактор

При правене на препоръки на новини следва да се отбележи, че интереса към тях е временен. Новините остаряват с времето, ставайки по-малко интересни за потребителя и поради тази причина това трябва да се отрази и в математическият модел за оценяването им. За адаптиране на фактор на забравяне в математическият апарат следва да се разгледа как е добавен той в най-известните онлайн новинарски източници като Google News⁴, Reddit⁵ and Hacker news⁶. За Google няма официална информация за метода за забравяне, който ползват, но в интернет пространството вече има официална информация за формулите които се използват за оценяване на интереса към различните новини за Reddit и Hacker News (Dover, 2008) (Salihefendic, 2010) (Salihefendic, 2010b). И двата сайта показват неперсонализирани препоръки базирани основно на глобалният интерес към различните новини. Reddit използва за оценяване на новините си следната формула:

$$\log_{10} \max(1, |U - D|) + \frac{(U - D)t_{post}}{45000} \quad (3-7)$$

Където с U обозначен броя на потребителите гласували, че харесват новината, а с D е обозначен броя на негативните гласове. t_{post} е времеви отпечатък на времето, в което е публикувана новината. При излизане на по-нови

⁴ <http://news.google.com/>

⁵ <http://reddit.com/>

⁶ <http://news.ycombinator.com/>

новини техният времеви отпечатък е по-голямо число и по този начин става автоматично намаляване на оценката на старите новини.

Hacker news ползват различна формула, като при тях се въвежда допълнителен наказателен член P , който се използва от редакторите на сайта за намаляване на оценките на някои видове. Във формулата на Hacker news се използва и променливата t_{now} , която е времеви отпечатък на текущия момент в който се изчислява оценката. Формулата, която използват е следната:

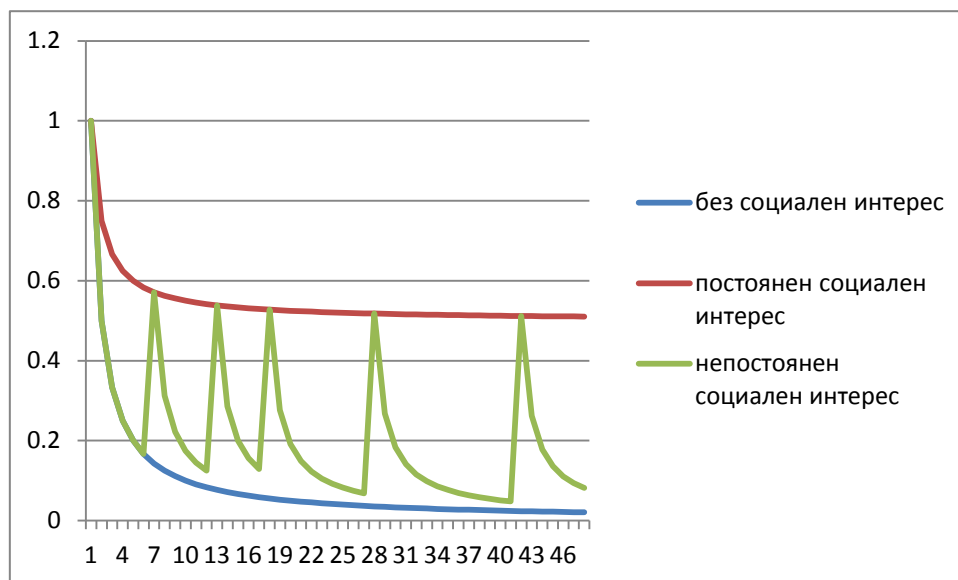
$$\frac{(U - D - 1)^{\alpha}}{(t_{now} - t_{post})^{\gamma}} P \quad (3-8)$$

В нашият случай за направа на оценка на различните новини в социалната мрежа първо използваме алгоритмите описани в предишните точки и след това прилагаме стратегия за забравяне по-подобен начин на забравянето внедрено в представените два новинарски сайта. Важно е да се отбележи, че в социалната мрежа, често новина е интересна заради добавен нов коментар или харесване към нея. Поради тази причина при остаряване на новината трябва да вземем в предвид едновременно кога е публикувана и кога последно някой от приятелите ни е взаимодействал с нея. Допълнително ако една новина е публикувана и в последствие никои не и е обърнал внимание искаме тази новина да остарява по-бързо отколкото новина, която е публикувана и в последствие постоянно харесвана и коментирана. Тези зависимости са отбелязани във формула 3-9, където $score$ е оценката за новината изчислена от модула за препоръчване, t_{post} , $t_{updated}$ и t_{now} са съответно кога новината е публикувана, кога последно е харесвана или коментирана и какъв е времеви отпечатък в момента.

$$score \left(\frac{1}{2(t_{now} - t_{post} + 1)} + \frac{1}{2(t_{now} - t_{updated} + 1)} \right) \quad (3-9)$$

Времето се представя във времеви отпечатък в часове. Това определя промяна на оценките на всеки час, като проявеният социален интерес към новината е решаващ за това колко бързо ще остарее. Фигура 12 показва различните начини за промяна на оценката с и без социален интерес към новината. Всяка новина започва с пълна оценка, която се редуцира с времето. Ако към новината няма социален интерес то тя след 6 часа вече е загубила 76% от оценката си, а след 12 часа – 92%, ако обаче потребителите харесват и коментират новината нейната оценка пада с 42% а след 12 с 47% като докато има интерес към новината нейната оценка не може да падне под 50% от първоначалната оценка.

При непостоянен социален интерес (новината е харесвана или коментирана от време на време) имаме постоянно остаряване на новината като при всяко взаимодействие на потребителите с нея нейната оценка се покачва над 50%, но след това отново почва бързо да губи от силата си.



Фигура 12. Времева промяна на оценката с различна степен на социален интерес.

Допълнително при изследване с реални потребители беше отбелязано, че новините не остаряват достатъчно бързо. Като основният забелязан проблем, е че Стари новини, към които е проявена инцидентно социална активност остават със сравнително висока оценка по-дълго време от желаното от потребителите. Този проблем бе решен с увеличаване на скоростта на остаряване на новината, когато към нея няма социална активност с модификация на формулата за остаряване до следния вариант:

$$score(\frac{1}{2(t_{now} - t_{post} + 1)} + \frac{1}{2(t_{now} - t_{updated} + 1)^2}) \quad (3-10)$$

3.5. Дискусия

Известната фраза „Каж ми кои са приятелите ти за да ти кажа ти кой си” важи в особено голяма сила за социалните мрежи. Това се обяснява с факта, че човек е социално същество и неговото мнение се влияе от мненията на неговите приятели. Това и възможността от предоставяне на обяснение на потребителя за направената препоръка определя очакването за преимущество на препоръките базирани на доверие в социалната мрежа. Доверието в препоръчващата система е

много важно за удовлетвореността на потребителя от направените му препоръки и поради тази причина очакването е, че система обясняваща препоръките си ще бъде възприета от потребителите по-добре отколкото система генерираща препоръки от черна кутия.

Допълнително следва да се обърне внимание на въпроса – Дали да препоръчваме само обекти, за които потребителя не знае или да препоръчваме и стари обекти, които той вече е харесал, но е възможно да иска да види отново? Отговора на този въпрос зависи от естеството на обектите, които се препоръчват. Ако препоръките се извършват върху обекти-новини, то много рядко някой потребител може да изяви интерес да му се препоръча новина, която вече е харесал преди години. В социалната мрежа обаче, интереса към обекта новина може да бъде запален отново, при наличие на коментар от приятел към тази новина. Отворен обаче остава въпроса дали след като новината е прочетена да остава в челото на потока с новини на потребителя или да губи от своята оценка за да може други непрочетени новини да заемат нейното място. От направеното проучване с потребители на социалната мрежа се вижда, че потребителите искат възможност за филтриране на прочетените новини. Това би предоставило възможност за решаване на проблема с предоставяне на избора, дали да се скрият прочетените новини или не, на потребителя. Отворен обаче остава въпросът какво означава, че новина е прочетена, понеже потребителя може да е погледнал една новина и тя да не му е била интересна в първият момент, но след като вече е харесана и коментирана от негови приятели тя да стане интересна за него, но да е вече скрита като прочетена новина за него.

Допълнителен отворен въпрос е как да се добавят нови новини в списъка с новини след като вече е бил показан на потребителя – дали да се запазва наредбата на новините до сега и новите да се добавят отгоре или да се вмъкват в списъка спрямо тяхната оценка за потребителя. Този въпрос няма еднозначно отговор. Ако подредбата се запазва и новите новини се слагат отгоре в списъка, то тогава потребителя ще знае какво е прочел и ще може лесно и бързо да прегледа новите обекти попаднали в потока му. Проблем при този подход ще бъде, че новите новини може да са доста по-неинтересни за потребителя, но те заемат челните места в потока му. Допълнително новина, която е надолу в списъка може да получи голяма социална активност и да стане по-интересна за него, но ако е

замразен потока да не се пренарежда няма да може да се покаже на по първо място в списъка му. Ако позволим списъка на потребителя да бъде пренареждан след като вече е бил представен на потребителя, може потребителя да се обърка като включваме нови новини измежду вече прочетените от него, но системата ще има възможността да покаже на потребителя точен списък с най-интересните за него обекти. Като компромисен вариант при реализирането на системата е избран подход, при който системата показва списък с препоръки, като социалната активност в тях се обновява постоянно, но техният ред се обновява само при явно указано желание от потребителя за зареждане на нови новини.

Глава 4. Система за препоръчване в социалната мрежа

Като част от дисертационният труд е изградена система⁷ за препоръчвана на информация в социалната мрежа Фейсбук. Системата използва програмният интерфейс на Фейсбук и достъпва данните за потребител и неговите приятели. Достъпа до системата се извършва през платформата на социалната мрежа, като по този начин се предоставя възможност за четене на потока от информация и използване на всички останали опции от социалната мрежа – известия за събития, съобщения от приятели и др. В следващите точки е разгледана в детайли архитектурата, функционалността и потребителският интерфейс на приложението.

4.1. Функционалност

След направен анализ на изискваната функционалност от потребителите (глава 2), достъпа до данни в социалната мрежа, различните свойства на препоръчващите системи и различните подходи за препоръчване на информация следната функционалност бе идентифицирана като необходима и възможна за реализиране:

1. Четец за новини
2. Търсеща машина за новини
3. Възможност за социална активност в приложението
4. Възможност за персонално отбелязване на интересни и неинтересни новини.
5. Автоматичен потребителски профил
6. Възможност за фина ръчна настройка на потребителския профил

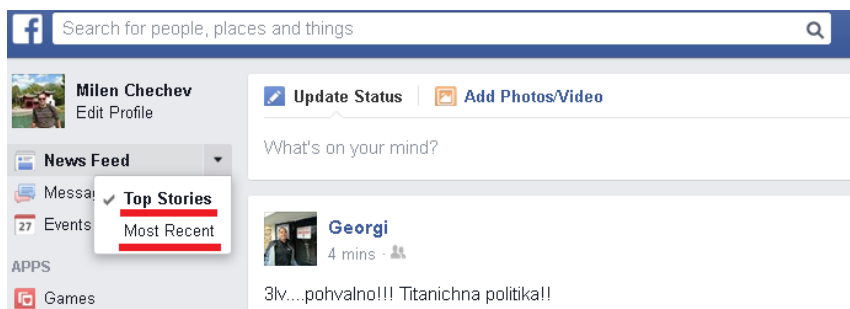
В следващите секции се дават детайли за всяка една от изброените функционалности.

4.1.1. Четец за новини

От направеното проучване сред потребителите става ясно, че основното време прекарано от тях в социалната мрежа е отделено за изчитане на потока им

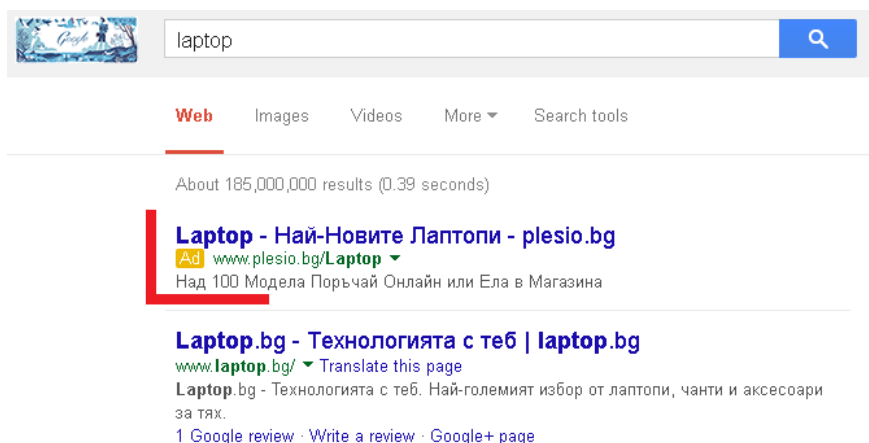
⁷ https://apps.facebook.com/recommended_stream

със социални новини. Фейсбук предлага два варианта за четене на потока с новини: „топ новини“ и „най-скорошни новини“ (Фигура 13), като Фейсбук строго препоръчва изгледа „топ новини“ и ако потребител прехвърли изгледа си на „най-скорошни новини“ постоянно е известяван да се прехвърли обратно на изгледа „топ новини“.

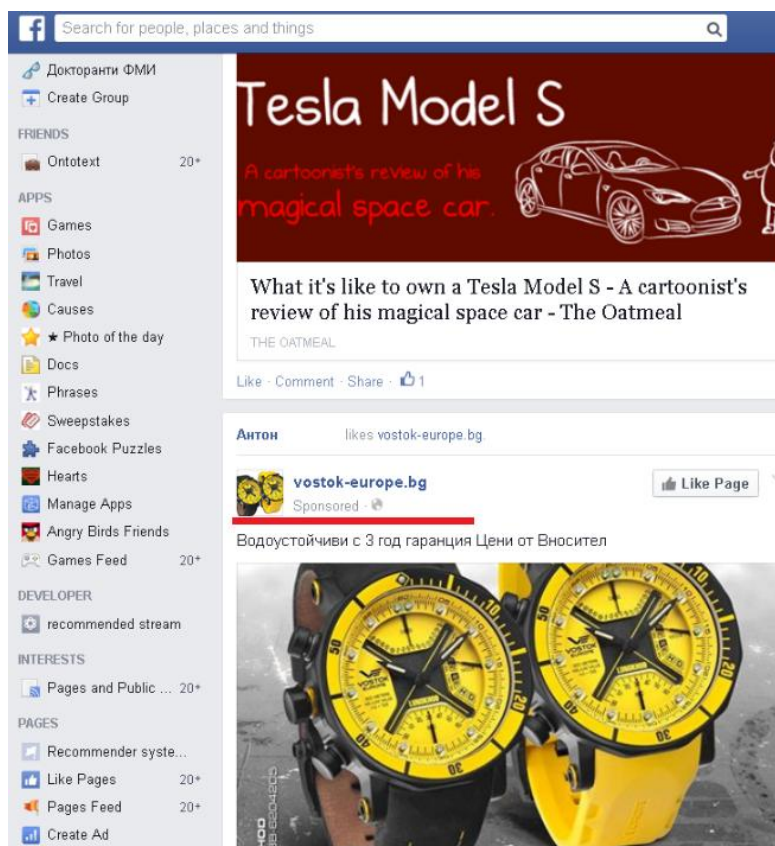


Фигура 13. Изгледи във Фейсбук: Топ-новини и най-скорошни новини

Потребителите нямат контрол върху подредбата на новините в изгледа „топ новини“. Фейсбук определя наредбата им, като дава възможност за показване на новина по-високо в потока на потребителите срещу заплащане. Това са така наречените спонсориран новини. Социалната мрежа смесва спонсорираните новини и социалните новини за да може спонсорираните да могат да получат по-добра видимост. Това се прави с цел да се повишат приходите от реклама на социалната мрежа давайки възможност за позициониране на реклами директно в потока с новини. За да бъдат повишено вниманието на потребителите към спонсорираните новини те не се разделят от останалите резултати, както например прави google при показване на резултати от търсене(Фигура 14). При Фейсбук първото рекламно съобщение се разполага на второ място в потока от новини, а следващите са разположение през 10 позиции едно от друго. Показването на реклама по-този начин максимизира на печалбата на социалната мрежа(10 процента от показаните съобщения са реклами), като същевременно прикрива рекламите в потока от новини (90 процента са истински новини).



Фигура 14. Показване на реклама в Google.



Фигура 15. Показване на реклама във Фейсбук.

В направено проучване сред 25 Фейсбук потребители се оказва, че около 20% процента от представените реклами са нерелевантни на потребителските интереси и потребителите нямат обяснение защо им се препоръчва подобен продукт. Потребителите виждат зависимост между тяхната активност в 45% от направените реклами, а останалите 35% си обясняват рекламата с това, че попадат в целевата група за рекламата от демографски характеристики.

Прекалено големият поток с новини и наличието на реклами започва да буди недоверие в потребителите към новините, които Фейсбук избира като

интересни за тях. Известно е, че социалните мрежи използват методи като търсене-използване (explore-exploit) за идентифициране на най-интересните новини за потребителите (Agarwal, 2013). Проблема с този подход за избиране на новини е, че социалната мрежа не предлага на потребителите си никакъв контрол върху филтрирането на информация и това може да доведе до филтрирането на интересни за тях новини. Фейсбук дори бива обвинена за това, че подвежда компании да популяризират продуктите си срещу заплащане, като в резултат вместо да получат повече взаимодействие с потребителите те получават по-малко (Veritasium, 2014).

Системата реализирана, като част от дисертацията разполага със функционалност за четене на потока с новини на потребителя. За извличане на новините се използва програмният интерфейс на Фейсбук, който връща само реалните новини без реклами и спонсорирани новини. Идеята е четецът за новини да спечели доверието на потребителя в показваната от него информация с помощта на прозрачна система за препоръчване на информация и предлагане на допълнителни услуги.

4.1.2. Търсеща машина за новини

Един от проблемите идентифициран в социалната мрежа Фейсбук е отсъствието на функционалност за търсене на новини. Липсата на тази функционалност ограничава потребителите и те изпитват затруднения да намерят интересна за тях публикация, която са прочели преди известно време. В рамките на изследването не са открити обективни пречки за реализиране на тази функционалност, като например социалната мрежа на Гугъл предлага тази услуга. Най-вероятната причина за отсъствието и е желанието на основателите на социалната мрежа техният продукт да се използва по-скоро като поток от информация вместо за хранилище за данни.

Реализираната система предлага на потребителите възможност за търсене в потока им с новини, като търсенето е ограничено до наличните данни за потока на потребителя от момента, в който се регистрира в приложението.

4.1.3. Възможност за социална активност в приложението

Приложението предлага възможности за социална активност – харесване, коментиране споделяне на новини, като добавянето на тези опции допринася за възможността то да бъде използвано като заместител на предлаганият от социалната мрежа четец за новини.

4.1.4. Възможност за персонално отбелязване на интересни и безинтересни новини.

Потребителското проучване показва, че опцията за „харесване” в някой се квалифицира като прекалено силна оценка за новина. След като потребител отбележи, че харесва новина, то неговото име се асоциира с нея и всеки, който има достъп до нея, вижда, че потребителя я е харесал. В някой случаи потребител предпочита да не отбелязва новина като интересна, понеже не иска публично да изрази това свое предпочитание. Това води до некоректно събрана информация от социалната мрежа относно интересните новини на потребителя.

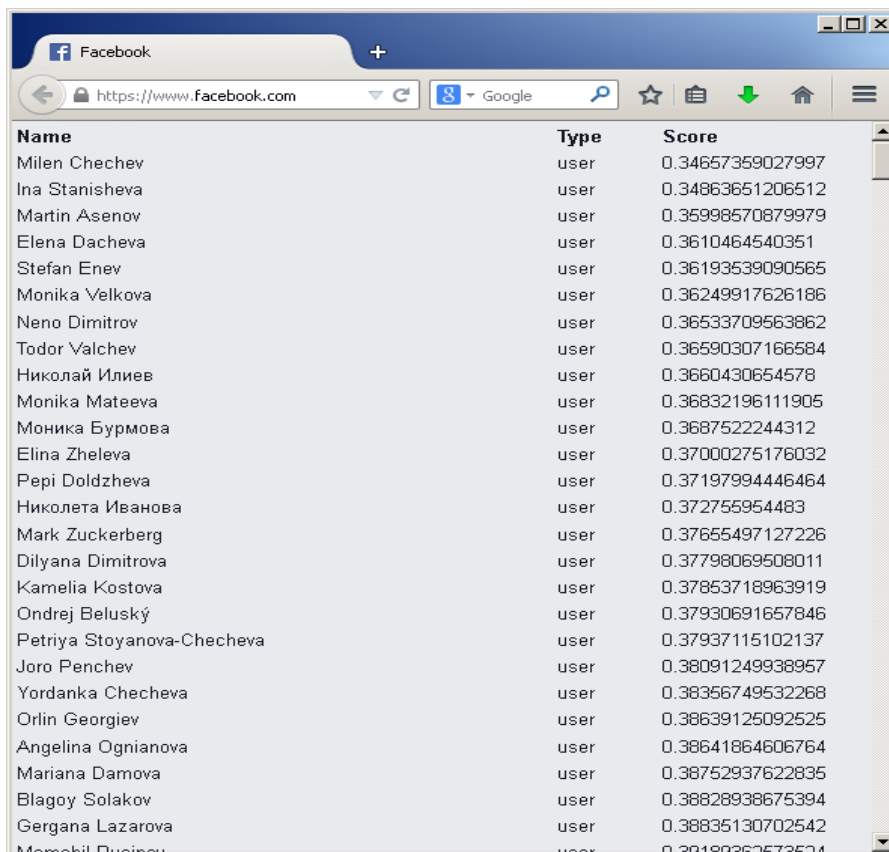
В предлаганата система се интегрира възможността да се отбележи новина като „интересна” или „неинтересна”. Този вид действие не е видимо за приятелите на потребителя, но допринася за по-добро разбиране на интересите на потребителя. При отбелязване на новина за интересна системата обръща допълнително внимание на съдържанието на новината.

За стимулиране на потребителите да използват тази функционалност се предлага опция за разглеждане на списъците с интересни и неинтересни новини, като по този начин потребителя може да използва списъка с отбелязаните си интересни новини за разглеждане по-късно.

4.1.5. Изграждане на автоматичен потребителски профил

През септември 2013 Фейсбук публикува новина за промени в потребителският поток с персонализиране на показваните реклами (Ge, 2013). Това персонализиране се базира на автоматично изграден потребителски профил въз основа на активността на потребителя в социалната мрежа. За съжаление потребителите нямат достъп и не могат да се разглеждат генерираните за тях потребителски профили. Ноември 2013 в интернет бе публикуван скрипт, с който този профил може да бъде извлечен неявно (Sreedharan, 2013). Публикуваният

скрипт извлича оценките, които Фейсбук е съставил за всички потребители с които е имал взаимодействие потребителят, като ги представя в табличен вид (Фигура 16). Резултатите показани на фигурата са генерирани от указаният по-горе скрипт и не са публично предоставени от Фейсбук.



Name	Type	Score
Milen Chechev	user	0.34657359027997
Ina Stanisheva	user	0.34863651206512
Martin Asenov	user	0.35998570879979
Elena Dacheva	user	0.3610464540351
Stefan Enev	user	0.36193539090565
Monika Velkova	user	0.36249917626186
Neno Dimitrov	user	0.36533709563862
Todor Valchev	user	0.36590307166584
Николай Илиев	user	0.3660430654578
Monika Mateeva	user	0.36832196111905
Моника Бурмова	user	0.3687522244312
Elina Zheleva	user	0.37000275176032
Pepi Doldzheva	user	0.37197994446464
Николета Иванова	user	0.372755954483
Mark Zuckerberg	user	0.37655497127226
Dilyana Dimitrova	user	0.37798069508011
Kamelia Kostova	user	0.37853718963919
Ondrej Beluský	user	0.37930691857846
Petriya Stoyanova-Checheva	user	0.37937115102137
Joro Penchev	user	0.38091249938957
Yordanka Checheva	user	0.38356749532268
Orlin Georgiev	user	0.38639125092525
Angelina Ognianova	user	0.38641864606764
Mariana Damova	user	0.38752937622835
Blagoy Solakov	user	0.38828938675394
Gergana Lazarova	user	0.38835130702542
Marek Bil	user	0.39190267572574

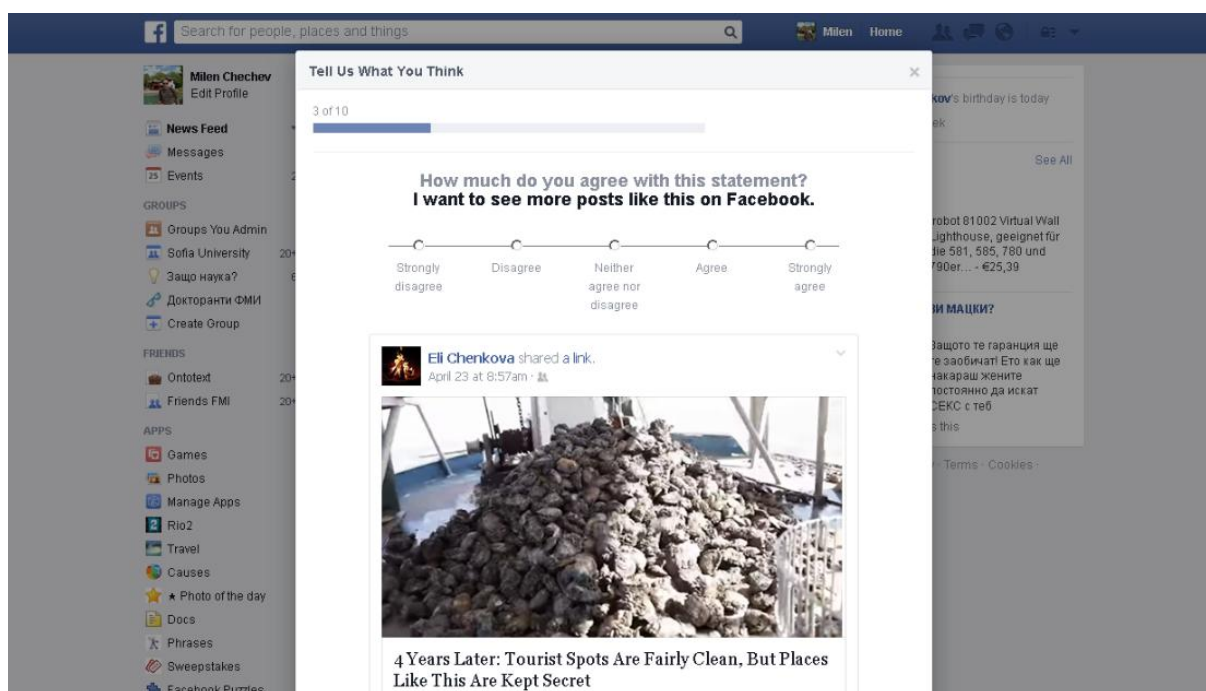
Фигура 16. Представяне на предпочитанията към потребители от скрипта на Sreedharan

Като част от дисертацията беше направено проучване между 10 Фейсбук потребителя за това доколко оценките, които се извличат от Фейсбук отговарят на начина, по който те биха наредили своите приятели. 80% от тях намериха поне един приятел измежду класираните в топ 10, който те не мислят, че им е толкова добър приятел и не се интересуват много от него. Допълнително разглеждайки наредените приятели на последните места повечето потребители намериха поне един приятел, които те не биха поставили измежду най-далечните си приятели. Не предоставянето на явна възможност за преглеждане на потребителския профил, поражда съмнение относно способността на социалната мрежа да определя от какво се интересува потребителя. При несправяне с този проблем социалната мрежа в претърпяла загуби свързани със спадане на лоялността на потребителите и загуба на част от тях.

В приложението разработено като част от дисертацията се добавя функционалност за разглеждане на генерираният от системата профил. По този начин системата предоставя възможност на потребителя да разбере защо получава точно такъв вид препоръки и той добива по-голямо доверие в нея.

4.1.6. Възможност за фина ръчна настройка на потребителския профил

Социалната мрежа не позволява на потребителите да укажат явно предпочитанията си за сортиране на техният поток с новини. Фейсбук събира главно неявна информация за активността на потребителя, като в началото на 2014 е забелязано и еднократно явно събиране на информация, чрез допитване до потребителя (Фигура 17). При това допитване се пита явно потребителя дали предложеното му съдържание е интересно. След попълване на анкетата социалната мрежа благодари на потребителите и скрива анкетата от тях, като в последствие те не могат, ако искат, да продължат с обучението на системата.



Фигура 17. Явно събиране на предпочитания за новини във Фейсбук.

В предлаганата система освен възможност за разглеждане на генерираният потребителски профил се предлага и възможност допълнително ръчно профилиране. Тази възможност е добавена като искана от голям процент от изследваните потребители в глава 2 и като важна за потребителите за да имат

възможност да повлияят на получаваните препоръки и да обучат системата, какво точно ги интересува.

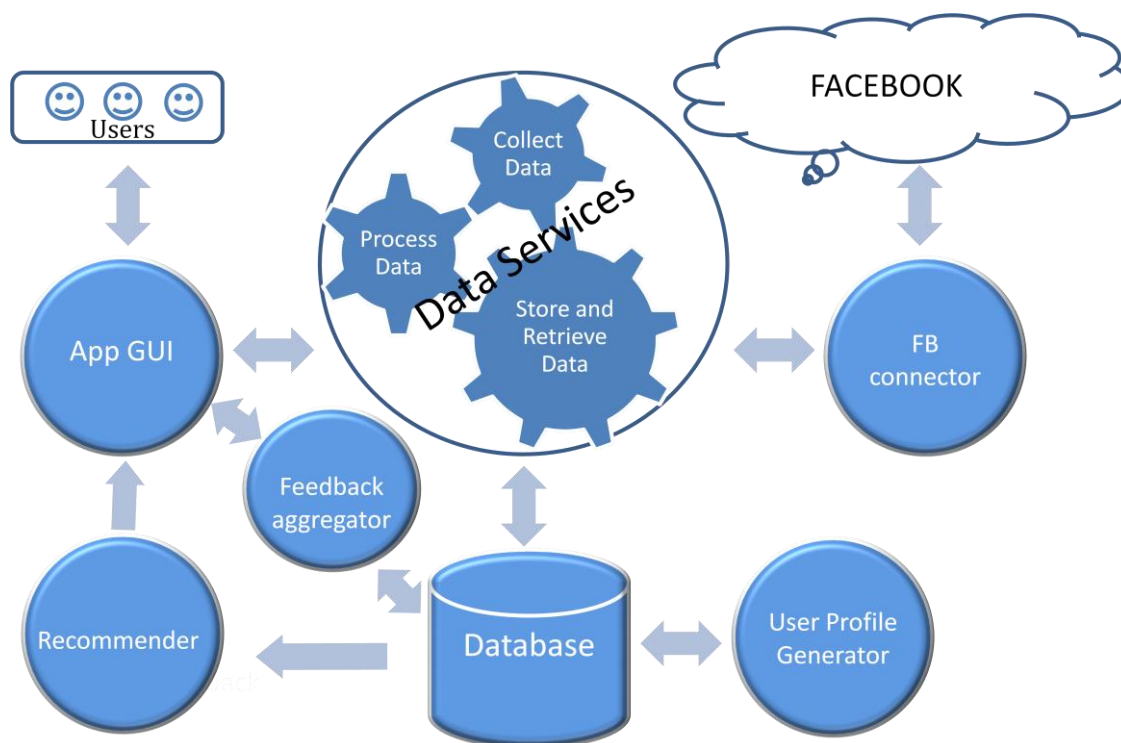
Възможностите за ръчна настройка включени в системата са възможност за пренареждане на реда на приятелите и възможност за абониране за ключови думи.

4.2. Архитектура на системата

Системата се състои от няколко основни компоненти-модули (Фигура 18).

- Графичен интерфейс – визуалната част от приложението предлагаща на потребителя различни изгледи на данните извлечени от Фейсбук.
- Модул Data Services в който са всички услуги за работа с данните извлечени от Фейсбук. Той се свързва с модула FBConnector и го използва за връзка със социалната мрежа и извличане на данните. Прави обработка на данните и ги съхранява в хранилището за данни.
- Модула FBConnector отговаря за връзката с Фейсбук, изпращане на заявки към социалната мрежа и получаване на данните
- Database е хранилището за съхранение на данни. Данните, които ни се предоставят от Фейсбук са в json формат и не са подходящи за съхранение в релационна база данни. Има необходимост от използване на нерелационна база данни, като в конкретният случай е използван MongoDB.
- Модула UserProfileGenerator използва данните събрани в хранилището с данни и изгражда потребителски профили. Създадените потребителски профили се съхраняват в хранилището с данни, където в последствие се извличат от модула Recommender за да послужат за направата на препоръки.
- Модула Recommender изготвя препоръки използвайки данните за обектите достъпни за потребителя и неговият профил.
- Модула FeedbackAggregator получава информация от потребителския интерфейс за активността на потребителя обработва я и я запомня в хранилището с данни. Потребителя може

да предостави явна или неявна обратна връзка. Явна обратна връзка предоставя когато променя потребителският си профил. Неявна обратна връзка получаваме от действията на потребителя в приложението – колко време е използвал приложението, колко обекти е разгледал, колко време се е задържал на тях, харесал ли е някой от тях и т.н.



Фигура 18. Архитектура на Фейсбук приложение за препоръчване на информация

4.3. Потребителски интерфейс

Приложението⁸ включва няколко екрана

- Въвеждащ екран, на който се обяснява основната функционалност на приложението и се дават добри практики за използването му. (Фигура 19). В началото потребителя дава права за достъп до потока с новини и в последствие има възможността да изгледа видео обясняващо функционалността на приложението докато приложението успее да изтегли достатъчно данни за да ми покаже

⁸ https://apps.facebook.com/recommended_stream

резултат. Обикновено зареждането на данни отнема около 1-2 минути, с което е съобразена и дължината на входящото видео.

Recommended Stream Application

Hello and welcome,

The Recommended Stream application is designed to help user's work and life at the social network Facebook.

It's main idea is to provide services to the users that will assist their everyday interaction with the social network.

The main services that the application is design to provide are:

- easy to use and configurable fully functional Facebook news feed reader.
- search functionality at the news feed.
- user profile that is automatically adjusted to your last activities, but can be manually changed
- option to manually or automatically re-rank your news feed to answer your preferences.

Current version provide just the basic functionality. Please, do not hesitate to report bugs or suggestions directly to the author. Contacts can be found at the about page.

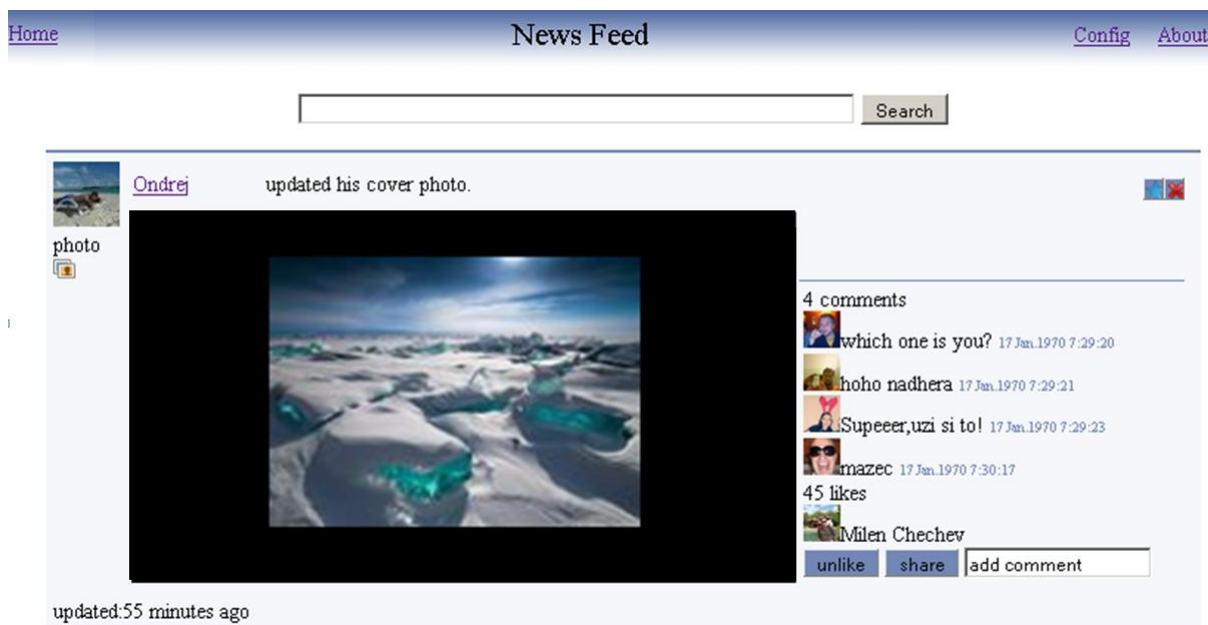
Video tutorial:



Loading your news feed, please wait...

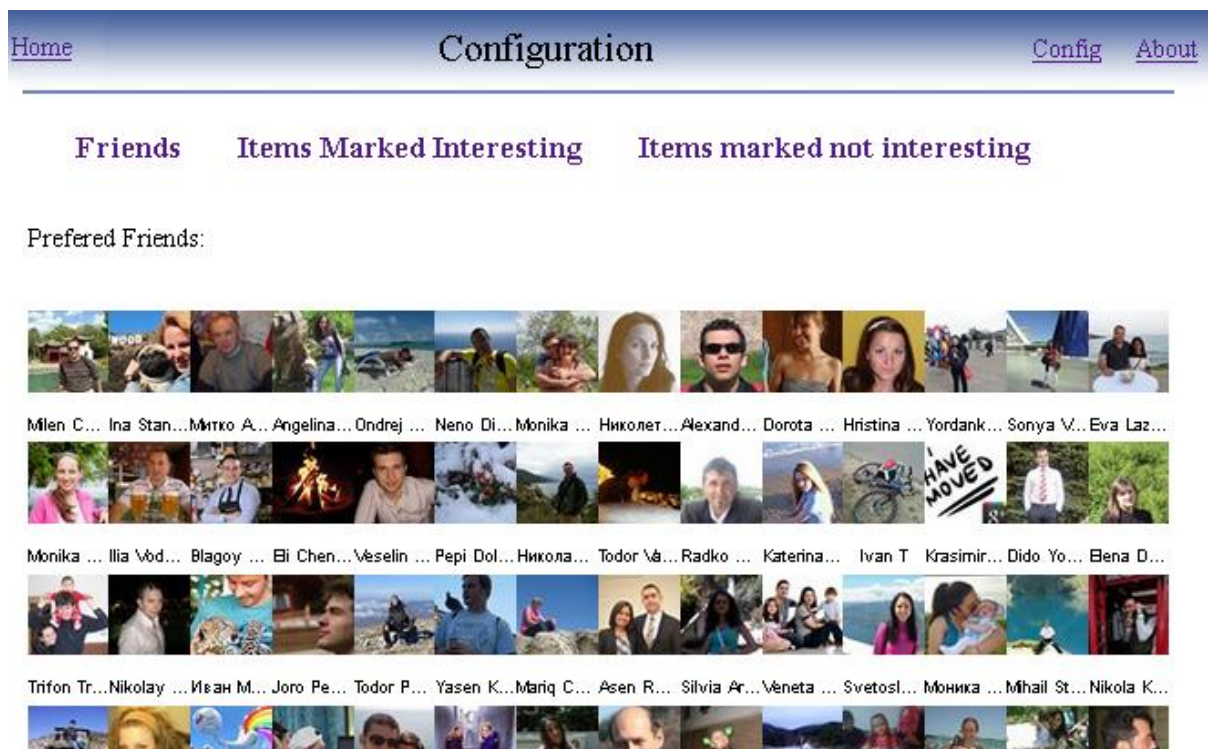
Фигура 19. Изглед от приложението за нов потребител.

- Страница за показване на новини, на която се показват новините публикувани от приятелите на потребителя (Фигура 20). За всяка новина потребителя вижда отбелязаните коментари и харесвания, като има възможността да „хареса“, „коментира“, „сподели“ или отбележи новината като „интересна“ или „неинтересна“.



Фигура 20. Основен изглед от приложението.

- Профил страница на която потребителя може да види какво знае системата за него и да направи фина настройка на детайлите(Фигура 21). Възможностите за промяна, които се предлагат на потребителя са: да промени позицията на приятел в списъка с приятелите, да разгледа списъка с харесани и нехаресани новини и да ги коригира изтривайки новина от тях.



Фигура 21. Изглед от страницата за потребителски профил

- Администраторска страница, на който администратор на приложението може да разгледа статистики за регистрираните потребителите, за тяхната активност, брой извлечени документи от техният поток и потока на техните приятели и т.н.

Глава 5. Софтуерни инструменти и реализация

В тази глава са разгледани детайли софтуерните инструменти използвани за реализацията на прототип на приложение препоръчващо информация в социалната мрежа. Секция 5.1 представя списък от най-известните инструменти използвани за реализация на препоръчващи системи и интерфейса за работа с две от най-популярните социални мрежи, а секция 5.2 представя детайли около конкретната реализация на прототипа.

5.1. Платформи и софтуерни инструменти

В тази точка ще разгледаме някой от най-известните софтуерни платформи за препоръчване предлагащи реализирани основните алгоритми за препоръки и възможности за разширяване.

5.1.1. Lens Kit

LensKit⁹ (Ekstrand, Ludwig, Kolb, & Riedl, 2011) е проект на изследователската група GroupLens, която е една от първите и най-известни групи за препоръчващи системи (Resnick, Iacovou, Suchak, Bergstrom, & Riedl, 1994). Системата е с отворен код и под GNU Public 2.1 лиценз. Предназначението на системата е да се използва с цел обучение и демонстрация, но би могла да се използва и в комерсиални проекти. Основните реализирани алгоритми във версия 2.0.5 са препоръки базирани на сътрудничество и близост между продукти (Deshpande & Karupis, 2004), препоръки базирани на сътрудничество и близост между потребители (Resnick, Iacovou, Suchak, Bergstrom, & Riedl, 1994), SlopeOne (Lemire & MacLachlan, 2007) и препоръчване с факторизиране на матрици по метода на Фънк (Funk, 2006)

5.1.2. Apache Mahout

Махаут (Apache Mahout¹⁰) е платформа за извличане на закономерности от данни, която набляга на дистрибутивния подход (Anil, Dunning, & Friedman, 2011). Системата е с отворен код с лиценз Apache v.2.0 и е използвана широко както за

⁹ <http://lenskit.grouplens.org/>

¹⁰ <https://mahout.apache.org/>

учебни така и за комерсиални цели. Версия 0.9(излязла февруари 2014) реализира различни разпределени алгоритми за класификация, клъстеризация и препоръчващи системи. В Махаут са включени всички алгоритми реализирани в LensKit(алгоритъма SlopeOne е премахнат от Махаут като неефективен във версия 0.8) и допълнително са реализирани алгоритмите за факторизиране на матрици по метода на редуващи се най-малки квадрати и теглова регуларизация (Zhou, Wilkinson, Schreiber, & Pan, 2008), факторизиране на матрици с паралелен стохастичено градиентно спускане (Takács, Pilászy, Németh, & Tikk, 2009) и факторизиране на матрици по метода SVD++ (Koren, 2008)

Махаут е най-популярната платформа за разпределени алгоритми за машинно самообучение, като е използван в множество комерсиални и академични проекти.

5.2. Достъп до данни в социалните мрежи

За експериментите извършени като част от дисертацията са използвани реални данни от две от най-популярните социални мрежи – Фейсбук и Туитър и двете предоставят програмен интерфейс за достъп, който ще бъде описан накратко в следващите подточки.

5.2.1. Facebook API

Фейсбук предлага два интерфейса за достъп до данните си: Graph Api¹¹ и FQL¹². При единият се получава информация чрез оказване на обект от социалният граф и/или връзка(релация) с него, а при FQL имаме възможност да изпълняваме заявки на опростен SQL подобен език. Двата подхода си имат своите предимства и недостатъци, но Graph Api е по-естествен за поддържане от социалната мрежа за това през май 2014 Фейсбук обяви, че поддръжката на FQL ще бъде спряна в следващите версии на програмния интерфейс.

Основните обекти, които могат да се извлекат от социалната мрежа чрез предоставеният интерфейс са:

/achievement	Обекти придобити от потребителите от Фейсбук игра
------------------------------	---

¹¹ <https://developers.facebook.com/docs/graph-api>

¹² <https://developers.facebook.com/docs/reference/fql>

/achievement-type	Описва типа на придобитият обект от Фейсбук игра.
/album	Фото албум създаден от потребителя. Може да се видят коментари към него или списък със снимките в него
/app	Обект за Фейсбук приложение. Към потребителите се асоциира такъв обект за всяко приложение, което добавят
/app-link-host	Връзка към сървър на който върви приложението добавено от потребителя.
/comment	Обект-коментар. Съдържа текста на коментар, датата на която е направен и харесванията на този коментар
/domain	Уеб домейна за Facebook Insights приложението
/event	Обект-събитие. Описва Фейсбук събитие
/friendlist	Списък с приятели. Всеки потребител може да групира приятелите си в различни списъци
/group	Фейсбук група за интереси. Всеки потребител може да членува в различни групи за интереси
/link	Обект-връзка. Показва хипервръзка споделена от Фейсбук потребител.
/message	Обект-съобщение. Този обект показва лично съобщение написано от един потребител към друг потребител.
/milestone	A Facebook Page milestone.
/object/comments	Този обект показва множеството от коментари към конкретен обект.
/object/likes	Този обект показва множеството от харесвания към конкретен обект
/object/sharedposts	Този обект показва множеството от споделяния на конкретен обект.
/object/insights	Този обект показва статистики за използването на обекта
/offer	Оферта публикувана от някоя страница
/page	Обект асоцииран с Фейсбук страница
/payment	Детайли за плащане направено през платформата на Фейсбук

/photo	Обект за снимка публикувана във Фейсбук
/place-tag	Отметка за потребител отбелязан на снимка, видео, новина, статус или хипервръзка.
/post	Новина (текстова публикация) публикувана във Фейсбук.
/review	Ревю за Фейсбук приложение.
/status	Обект за статус публикуван от потребител.
/test-user	Обект за тестов потребител създаден от Фейсбук приложение.
/thread	Нишка със съобщения във Фейсбук.
/user	Обект за потребител на Фейсбук
/video	Обект за видео публикувано във Фейсбук

Таблица 1. Таблица с различните видове обекти предоставяни от програмния интерфейс на Фейсбук

За достъпа до изброените обекти потребителя трябва да делегира разширени права за приложението, като Фейсбук предупреждава потребителя за това до каква информация ще има достъп приложението.

За връзка с програмния интерфейс на Фейсбук се достъпват REST услуги отворени от социалната платформа като входа и изхода им е организиран в json формат (Фигура 22)

```
{
  "data": [
    {
      "id": "228891633872530\_660010834093939",
      "from": {
        "category": "Tours/sightseeing",
        "name": "София (Sofia)",
        "id": "228891633872530"
      },
      "message": "Красив кадър!\n\nСнимка: Константин Иванов",
      "picture": "https://fbcdn-photos-d-a.akamaihd.net/hphotos-ak-xfp1/...",
      "link": "https://www.facebook.com/Sofia.City.Bulgaria/photos/a.2926087541...",
      "icon": "https://fbstatic-a.akamaihd.net/rsrc.php/v2/yz/r/StEh3RhPvjk.gif",
      "privacy": {
```


- friends/ids – връща идентификаторите на подмножество от приятелите на даден потребител.
- users/lookup – връща информация за публичните профили на до 100 потребителя,

Голямото предимство на Туитър пред Фейсбук е, че в него всичко публикувано е публично. Тази политика улеснява значително достъпа до информация и е определяща за по широкото използване на Туитър за редица експерименти.

Въпреки по-лесното извличане на данни от Туитър за експериментите описани в дисертацията са използвани главно данни от социалната мрежа Фейсбук заради по-голямата и популярност и по-сложните взаимоотношения и връзки между потребителите в нея.

5.3. Прототип

За изграждане на прототип на препоръчваща система на съдържание във социалната мрежа Фейсбук е използвана архитектурата, функционалността и потребителският интерфейс описани в глава 4.

За реализацията на потребителските услуги е използвана Java EE, за потребителски интерфейс е използван HTML и Javascript, за съхраняване на данни нерелационната база данни – MongoDB, за осъществяване на търсене Apache Solr, за часта от експериментите свързани с препоръки базирани на сътрудничество е използвана и платформата Apache Mahout.

За улесняване на достъпа до интерфейса на социалната мрежа Фейсбук е използвана библиотеката RestFB ¹⁵, като по-този начин се опростява комуникацията със социалната мрежа и обработката на възникнали грешки.

Програмният код на прототипа е отворен и предоставен за публично ползване в хранилище за контролна версиите базирано на системата Git – GitHub връзка: <https://github.com/milenchechev/RecommendedStream>

За разработка на приложението са използвани две различни Фейсбук приложения – за тестване (http://apps.facebook.com/recommended_stream_d) и за

¹⁵ <https://github.com/revetkn/restfb>

реална употреба (http://apps.facebook.com/recommended_stream).

Конфигурацията за това дали приложението се стартира за тестване/разработка или публично е изнесена в конфигурацията на сървъра, което позволява лесна разработка и автоматично качване на едно и също приложение на публична и тестова среда.

При работа на прототипа с реални потребители бе забелязан, проблем със бързодействието на индексирането на документи от търсещата машина Apache Solr, което би могло да доведе до проблеми с разширяемостта на приложението при работа с много потребители. За решението на този проблем в бъдеще се планира да се замени Solr с Elasticsearch.

Не са забелязани проблеми със останалите, компоненти, като са направени тестове с 300 тестови потребители във Фейсбук.

Глава 6. Оценяване

Системата-прототип е направена да работи с препоръки базирани на доверие, предоставяйки по този начин на потребителя препоръки за публикации от неговите приятели подредени спрямо предпочитанията му към тях. За оценяване на системата са направени следните експерименти:

- Офлайн оценяване на точността на препоръчване на информация базирано на съдържание, сътрудничество, доверие и хибриден подход.
- Онлайн оценяване на системата с реални потребители следейки тяхното взаимодействие със прототипа с неявна и явна обратна връзка.

6.1. Събиране на данни

Данните за оценяване на системата са събрани от социалната мрежа Фейсбук. Фейсбук е социална мрежа от затворен тип, което не позволява да се събират данни без потребителите явно да дадат права за това. Системата е тествана с 25 реални Фейсбук потребители, повече от които студенти в магистърски програми във Факултета по математика и информатика на Софийски университет.

Приложението започва да събира данни за потребител при добавянето му към профила му и делегирането на необходимите права. Първоначалното събирането на информация от потребител отнема около две минути, като се събират данни за текущият поток от новини на потребителя и потока от публикации за всички негови приятели (въвежда се ограничение от максимум 500 приятели за първоначалното зареждане).

След първоначалното зареждане на данните се генерира потребителски профил въз основа на получената информация, предоставя се на потребителя възможност да работи с приложението, а то продължава с изтеглянето на информация за последните 1000 харесани обекта от потребителя и публикуваните албуми със снимки от приятелите на потребителя.

Статистики за събрани данни са представени в Таблица 2

Брой регистрирани потребители	25
Брой на приятелите	6548
Публикации от поток с новини	20710
Публикации публикувани от приятели	853366
Харесани обекти	19212

Таблица 2. Статистики за събраните данни.

6.2. Офлайн оценяване на системата

Целта на офлайн оценяването е да провери предположението, че препоръките базирани на доверие намират релевантни публикации за потребителя и са сравними, а в някои аспекти и по-добри от препоръките базирани на съдържание или сътрудничество. При тези изследвания не е отчетено времето, в което са публикувани обектите и не е използван фактор за остаряване.

За проведените експерименти е използвана 10 групов (10-fold) кросвалидация. Като сравняваните подходи са препоръки базирани на съдържание, препоръки базирани на сътрудничество, препоръки базирани на доверие и хибриден подход използващ паралелна архитектура. При комбиниране на подходите в хибриден модел се прилага обучение на теглата за комбиниране с градиентно спускане.

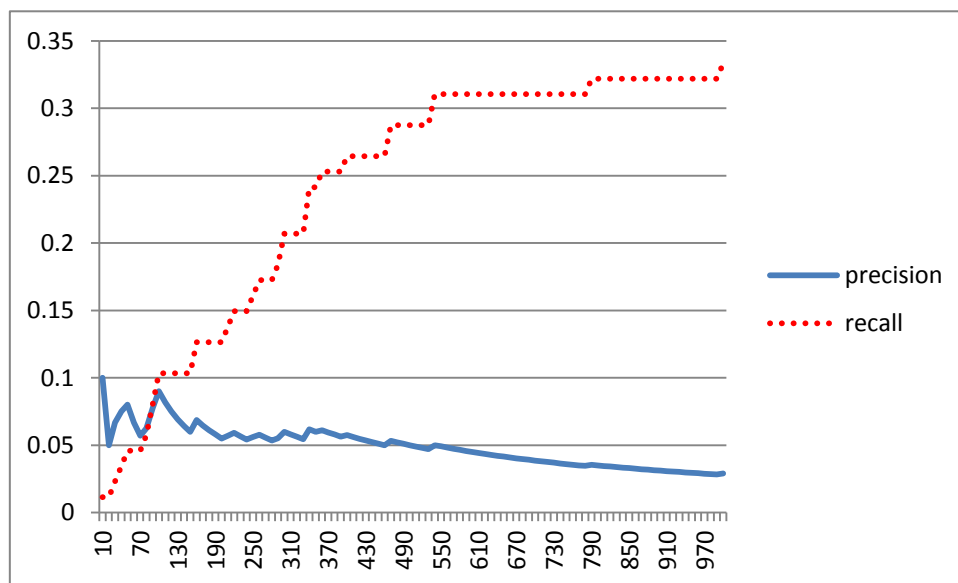
6.2.1. Изследване за правилният брой резултати за препоръчване

За всеки от изследваните подходи са направени експерименти за определянето на точността и откриването при използване различен брой резултати от препоръчващата система. Поради наличието на повече от един потребител и съответно повече от един списък с резултати за визуализиране средно аритметичната стойност от всички резултати, т.е. използваме формулите:

$$mean_precision@N = \frac{\sum_{for\ all\ users} precision@N}{\#number\ of\ users} \quad (5-1)$$

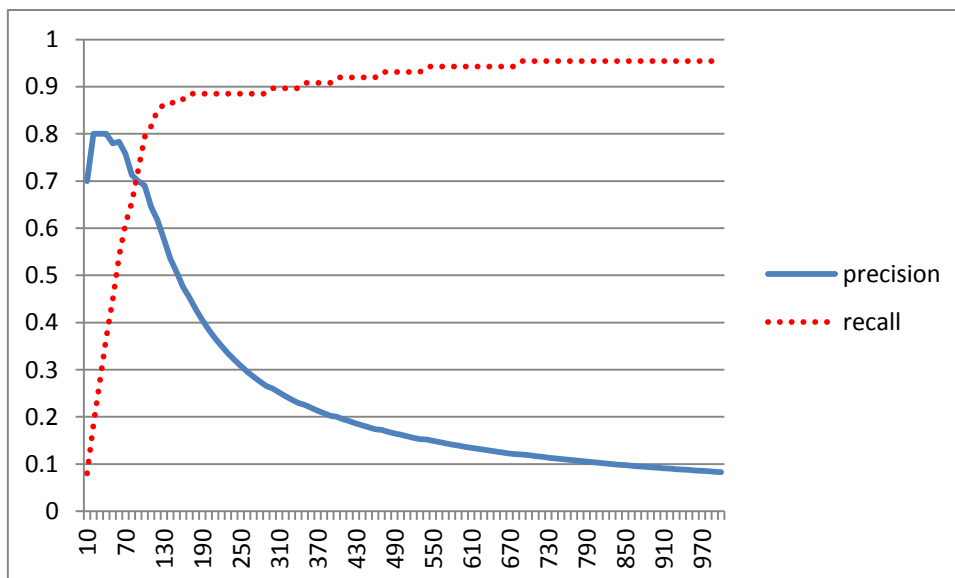
$$mean_recall@N = \frac{\sum_{for\ all\ users} recall@N}{\#number\ of\ users} \quad (5-2)$$

На Фигура 23 е показано точността и откриването при препоръчване базирано на съдържание. Абцисната ос показва броя на препоръчаните примери, а ординатната стойността на точността и откриването. При препоръчване на около 100 примера имаме стойност от 0.1 едновременно за точността и откриването. Забелязва се, че броят откриването достига максимум 0.33 при препоръчани 1000 обекта на потребителя.



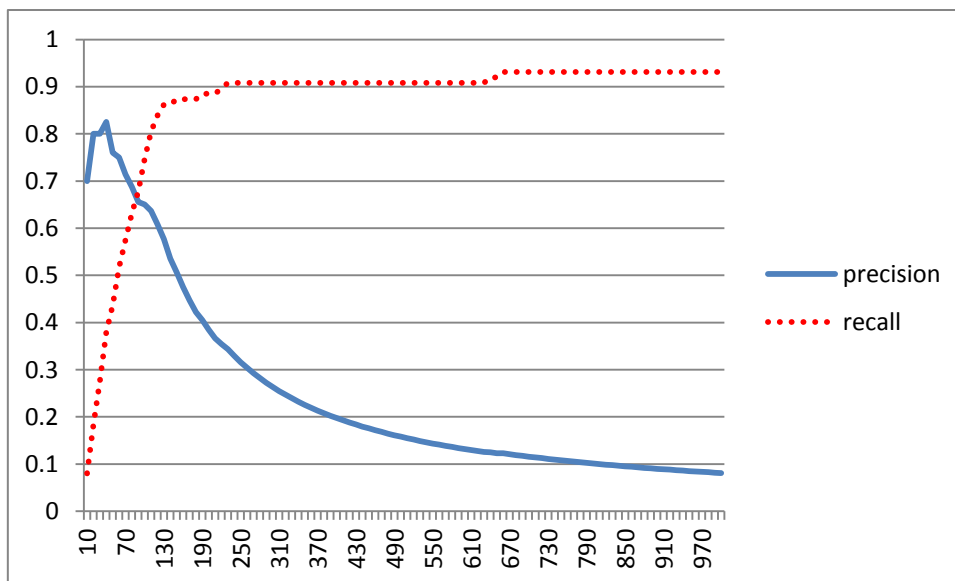
Фигура 23. Точност и откриване с препоръки базирани на съдържание.

Резултатите от препоръките базирани на сътрудничество са показани на Фигура 24. Забелязва се, че точността е значително по-висока от тази на препоръките базирани на съдържание като при препоръчване на 20 обекта имаме точност от 0.8, а при препоръчване на 100 обекта имаме откриване 0.8.



Фигура 24. Точност и откриване на препоръки базирани на сътрудничество.

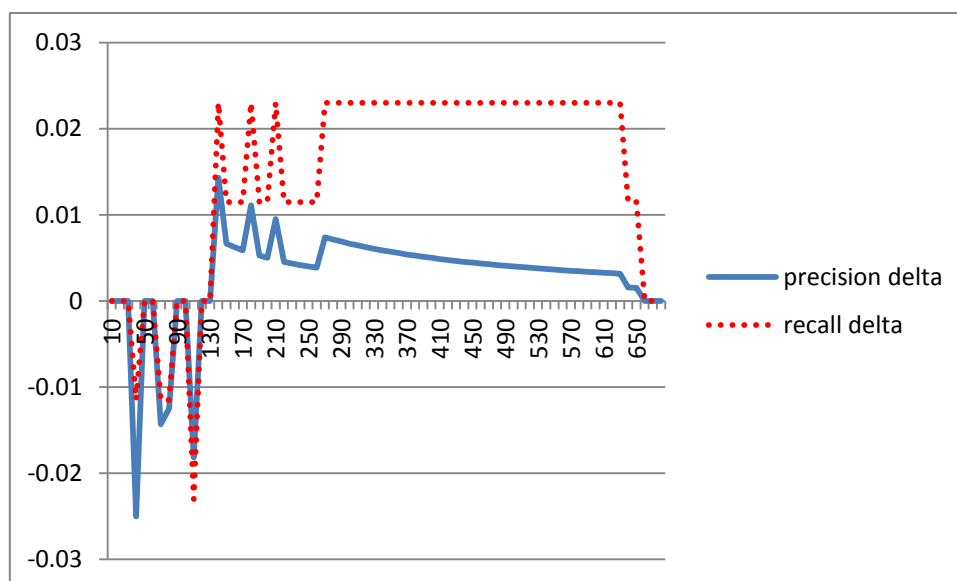
Следващата фигура показва точността и откриването при използването на доверие. При препоръките базирани на доверие се наблюдава отново високи стойности на точност и откриване, като се забелязват някои разлики с препоръките, базирани на сътрудничество. При препоръчани 40 обекта препоръките базирани на доверие имат по-висока точност от тази на препоръките базирани на сътрудничество – 0.825 срещу 0.8. Препоръките базирани на сътрудничество показват обаче по-висока точност при разглеждане на първите 70 препоръки – 0.757 срещу 0.714. Това поведение на по-силни препоръки в началото на списъка и не чак толкова силни следващи резултати е очакван резултат понеже при препоръките базирани на доверие потребителя получава с приоритет публикациите на най-добрите си приятели, но след като те свършат публикациите на приятелите с не чак толкова силно доверие в тях може да се окаже по-слаби от публикации на непознати хора близки до предпочитанията на потребителя.



Фигура 25. Точност и откриване на препоръки базирани на доверие.

При изследването на хибридният подход не са наблюдавани големи разлики с резултатите от подхода базиран на доверие, поради което за по-добра визуализация на Фигура 26 представяме разликата между метриците на подхода базиран на доверие и хибридният подход. Вероятно липсата на подобрение в резултатите на хибридният алгоритъм се дължи на използването на едни и същи данни и от всички комбинирани подходи, което води до извличането на подобни закономерности от тях и не допринася за обогатяване на резултатите при комбинирането на различните подходи.

В резултатите показани от хибридният подход прави впечатление, че точността на хибридният подход се движи около тази на препоръчването базирано на доверие, като по значително подобрение на резултатите се наблюдава чак след показването на 130 резултата.



Фигура 26. Оценяване на точността на хибридният подход.

Въз основа на показаните фигури за изменението на точността и откриването е направено заключение, че показването на около 100 новини на потребителя би допринесло за допустими стойности за едновременно точността (при подхода за препоръчване базиран на доверие тя е 0.65) и на откриването (при подхода базиран на доверие то е 0.74).

Относно избора на алгоритъм за препоръчване точността на препоръчването базирано на доверие е сравнима с точността на препоръчването базирано на сътрудничество при препоръчване на 100 новини. Точността на препоръките базирани на доверие в първата половина на резултатите е по висока от тази на препоръките базирани на сътрудничество, което определя препоръките базирани на доверие като по-силен кандидат за препоръчващ алгоритъм отколкото препоръките базирани на сътрудничество. Допълнителната възможност да се обясни на потребителя, че препоръките са му направени понеже публикацията е харесана и коментирана от негови доверени приятели окончателно мотивира избора на препоръките базирани на доверие като най-подходящ алгоритъм за препоръчване на съдържание в социалната мрежа.

6.2.2. Оценяване на подредбата на списъка с препоръки

За оценяване на подредбата на обектите в списъка с препоръки са използвани и метриките Mean Average Precision и nDCG. nDCG е метрика, която се

измерва върху само една последователност от резултати за това за измерването и обобщено върху всички потребители ще измерваме:

$$\text{mean_nDCG} = \frac{\sum_{\text{for all users}} \text{nDCG}}{\text{\#number of users}} \quad (5-3)$$

При пресмятане на nDCG използваме следната формула за обезценена кумулативна печалба

$$\text{DCG} = \text{rel}_1 + \sum_{i=2}^{100} \frac{\text{rel}_i}{\log_2(i)} \quad (1-31)$$

Където rel_i е равно на 1 ако обекта е харесан от потребителя и 0 ако не е харесан.

Mean Average Precision го пресмятаме използвайки различни потребители вместо различни заявки. Формулата е следната

$$\text{MAP} = \frac{\sum_{\text{for all users}} \text{AveP}(u)}{\text{\#number of users}} \quad (5-4)$$

Където $\text{AveP} = \frac{\sum_{k=1}^n P@k * \text{rel}(k)}{\text{\#relevant docs}}$, $\text{rel}(k)$ е 1 ако обекта на позиция k е харесан от потребителя и 0 в противен случай.

Таблица 3 показва изчислените стойности за MAP и Mean nDCG за различните алгоритми с използването на първите 100 препоръчани резултати за всеки потребител.

	MAP	Mean nDCG
Content based	0.113	0.540
CF	0.739	0.860
Trust	0.733	0.868
Hybrid	0.743	0.870

Таблица 3. Сравнение на подредбата на резултатите с различните подходи.

Резултатите отново показват, че препоръките базирани на съдържание не са подходящи за препоръчване на новини в социалната мрежа. Разглеждайки резултатите за средна аритметична точност (MAP) виждаме, че хибридната система дава малко по-голяма точност от препоръките базирани на сътрудничество, но от препоръките базирани на доверие е по-добра с една стотна. Обяснението за това е, че и при хибридният подход и при препоръките базирани на сътрудничество имаме по-висока точност на резултатите във втората половина на списъка, което се оказва достатъчно за това увеличение. При метриката нормализирана обезценена кумулативна печалба обаче наредбата в

първата половина има по-голяма сила, като по този начин хибридният подход е по-добър само с 2 хилядни, а препоръките, базирани на сътрудничество, дават по-лош резултат от препоръките, базирани на доверие.

6.3. Онлайн оценяване на системата

Като част от дисертацията е направено и онлайн оценяване на системата, в което взеха участие 32 участници. Всички те присъстваха на демонстрация на функционалността на системата и обяснение на подходите за препоръчване използвани в системата. След това всеки един от тях самостоятелно използва системата и в последствие отговори и на оценяваща анкета за приложението. Основната цел на това изследване е да се оцени доверието на потребителите към система за препоръки. Измерва се броят на прегледаните новини, харесаните новини, коментарите, новините отбелязани за интересни и не-интересни. След това явно се пита потребителя относно неговото мнение за препоръчващата система. Таблица 4 представя въпросите и дадените отговори в проценти. Където за въпросите е използвана ликертова скала със възможни отговори: 1=Твърдо не; 2=По-скоро Не; 3=Колебая се; 4=По-скоро да; 5=Да.

	1	2	3	4	5
Оценете от 1 до 5 подредбата на новините в приложението?	0.0%	6.3%	46.9%	31.3%	15.6%
Разбирате ли как работи препоръчващата система?	0.0%	0.0%	9.4%	62.5%	31.3%
Имате ли доверие в подхода който системата използва за препоръчване?	0.0%	0.0%	12.5%	37.5%	50.0%
Бихте ли искали да можете да нагласяте подредбата във оригиналния поток във Фейсбук?	0.0%	0.0%	15.6%	37.5%	46.9%
Разбирате ли подхода за остаряване на новини използван в приложението ?	0.0%	6.3%	28.1%	0.0%	65.6%
Този подход за остаряване съвпада ли с вашите предпочитания?	0.0%	0.0%	6.3%	78.1%	15.6%
Харесва ли ви функционалността за търсене и предоставените резултати?	0.0%	3.1%	0.0%	81.3%	15.6%
Харесва ли ви възможността да промените наредбата	0.0%	0.0%	0.0%	25.0%	75.0%

на приятелите си в потребителския профил?					
Използвахте ли възможността за промяна на наредбата на приятелите в приложението?	46.9%	0.0%	0.0%	0.0%	53.1%
Бихте ли искали подобна функционалност(да подреждате приятелите си) във Фейсбук?	9.4%	18.8%	0.0%	34.4%	37.5%
Харесва ли ви възможността да отбележите публикация за интересна без да слагате публично харесване "like"?	0.0%	0.0%	6.3%	25.0%	68.8%
Бихте ли използвали тази функционалност за запомняне на любими публикации?	0.0%	6.3%	12.5%	31.3%	50.0%
Харесва ли ви възможността да отбележите новина за неинтересна?	0.0%	0.0%	12.5%	18.8%	68.8%
Бихте ли използвали тази възможност за намаляне на приятелството с потребител или предпочитате само новината да се скрива?	9.4%	18.8%	31.3%	9.4%	31.3%
Каква е цялостната ви оценка за приложението?	0.0%	0.0%	12.5%	0.0%	87.5%

Таблица 4. Резултати от потребителска анкета за реализираното приложение.

Резултатите от анкетата показват, че потребителите като цяло имат доверие в препоръките дадени им от системата, като 87.5% са отговорили положително на този въпрос. Подобен е и процента на потребителите, които биха искали да имат възможност за подобна наредба на новините и в оригиналната страница на Фейсбук.

Относно остаряването на новините 93.2% са отговорили, че то съвпада с предпочитанията им, като прави впечатление, че голяма част от тях не са съвсем сигурни, което се вижда и от процента участници, които разбират метода за остаряване на новини – 65%. Промяната на потребителския профил е единствената функционалност, която получава 100% положителна оценка, въпреки че само 53% от потребителите са изпробвали възможността за ръчно калибриране на тази подредба. Това е показателно за това, че потребителите не искат да отделят много време за създаване на потребителски профил, но имат желание да знаят, какво знае за тях социалната мрежа. Допълнително това твърдение се затвърждава от сравнително високият процент на потребители,

които не искат подобна функционалност за калибриране на приятелството в оригиналната версия на социалната мрежа – 28.2%. 93.2 % харесват възможността да отбележат новина като интересна, като 81% от потребителите биха използвали тази функционалност за да си направят личен списък с любими публикации. Отбелязването на новина като неинтересна също е доста харесвана функционалност – 87.5% от потребителите са я харесали, но само 40% биха искали това отбелязване да коригира приятелството им с автора публикувал новината. Цялостната оценка за приложението е положителна като 87% от потребителите са отговорили, че го харесват.

Заклучение

В представената дисертация е разгледан проблема за препоръчване на информация в социалните мрежи. Проблема е актуален и е породен от нарастващата популярност на социалните мрежи, увеличаване на съдържанието в нея и затрупването на потребителя с информация. За решаването на проблема е направен обзор на различните подходи препоръчващи системи, проучено е поведението и нуждите на потребителите, изследвани са различни видове препоръчващи алгоритми и е изградено приложение препоръчващо новини в социалната мрежа Фейсбук.

При изготвянето на подредбата на новините се използва неявно извлечени тегла, за изчислението на които се използва социалната активност на потребител към неговите приятели. Формулата и подхода за избиране на параметрите и са авторска разработка, като направеното оценяване показва добри резултати на препоръките генерирани с използване на извлеченото неявно доверие. Като част от дисертацията е изследван и проблема за остаряване на новините. Предложени са две формули, като е предпочетена формулата с по-бързо остаряване на новина при ниска социална активност към нея.

Направеното оценяване на препоръчващите подходи показва, че подхода базиран на доверие има предимство пред останалите подходи, поради допълнителното предимството за възможност за обяснение на направените препоръки към потребителя, възможността за изграждане на явен потребителски профил и фина ръчна настройка на потребителския профил от потребителя.

Направено е и оценяване на удовлетвореността на потребителите от реализираното, като част от дисертацията, приложение за препоръчване на новини в социалната мрежа. Резултатите показват, че потребителите разбират направените препоръки и се доверяват на системата. Харесват възможността за разглеждане и промяна на потребителският си профил и за алтернативно разглеждане и търсене в новините от техният поток.

Перспективи за бъдещо развитие

Перспективите за бъдещото развитие на проекта е да бъде разширен и популяризиран като изследователски проект с отворен код. Прототипа може да се ползва като основа за изграждане на по-широкообхватно приложение, което да продължи изследването в някой от тези посоки:

- Популяризиране на приложението в социалната мрежа
- Изследване на допълнителни свойства за потребителския профил.
- Групиране на новините на потребителя по теми и динамично намиране на клъстери с новини.
- Акумулиране на информация от повече от една социална мрежа и евентуално други източници на информация – rss feed, blogs, forums
- Различно доверие към приятелите по-различни теми – от един приятел може да са ни интересни вицове, но не и новините, докато от друг може да е точно обратното.
- Различни настроения на потребителя – в различно настроение или време от деня, потребителя може да се интересува от различни неща. Да се изследва възможността да се поддържат повече от един профил за потребителя за различните му предпочитания.

Научни приноси

- Разработен е метод за оценяване на доверието на потребителите към техните приятели в социалната мрежа.
- Разработен е метод за състаряване на новините в социалната мрежа отчитащ времето на публикуване на новината и социалната активност към нея.

Научно-приложни приноси

- Направен е подробен анализ на проблемите на съществуващата система за препоръчване на новини във Фейсбук.

- Проектирана е препоръчваща система, удовлетворяваща нуждите на потребителите в социалната мрежа.
- Направено е сравнение между различни видове препоръчващи подходи в социалната мрежа.

Приложни приноси

- Разработване на приложение с отворен код за социалната мрежа Фейсбук, събиращо информация за регистрирани потребители и предоставящо им алтернатива, за разглеждане на потока от информация.
- Реализация на иновативен начин за представяне на потребителския профил на потребител и възможност за редактирането му.
- Добавяне на иновативна възможност в реализираният прототип за разширена явна обратна връзка. Тази обратна връзка е възможна чрез използването на маркерите за „интересна”/”неинтересна” новина.

Публикации

Публикации на Милен Чечев, свързани с тематика на дисертацията:

1. *Milen Chechev, Ivan Koychev. Social News Feed Recommender, AIMS'2014, September 2014 (пруга за печат)*
2. *Milen Chechev, Ivan Koychev. Social Recommendation Based on Content and Trust, 2014, Седма Национална конференция „Образованието и изследванията в информационното общество”, Май 2014*
3. *Milen Chechev, Ivan Koychev. Recommendations in Social Networks: an Extra Feature or an Essential Need, In Proceedings of MIE'2013, Sofia, Bulgaria, September 2013*
4. *Milen Chechev, Petko Georgiev. A multi-view content-based user recommendation scheme for following users in Twitter, In: Proceedings of SocInfo'2012, Lausanne Switzerland, December 2012, p. 434-447*
5. *Milen Chechev. Recommender approaches in the social network, Computer Science and Technologies Journal, Technical University, Varna, 2012*

6. *Milen Chechev, "Recommender systems challenges.KDD Cup 2011." In Proceedings of Days of Science, Veliko Turnovo, Bulgaria, May 2011*
7. *Pavel Boychev, Alexander Grigorov, Luigi Sarti, Atanas Georgiev, Krasen Stefanov, Milen Chechev, "Recommender systems and repository search: The SHARE.TEC proposal", "Sofia University Journal of E-learning", 2010*
8. *Milen Chechev, Ivan Koychev. Social Reading and Book Sharing Environment. Proc. of Int. Conference S3T'10,2010*

Доклади

1. *Milen Chechev, "Mood-based Recommender for Social Network", FET EYE Lab Surfing workshop, Солун, Гърция, Март 2014*

Декларация за оригиналност

Декларирам, че представената във връзка с провеждането на процедура за придобиване на образователната и научна степен „доктор“ в Софийски университет “Св. Климент Охридски“ дисертация на тема: “Система за препоръчване на съдържание в социалната мрежа “ е мой труд.

Цитиранията на всички източници на информация, текст, илюстрации, таблици, изображения и други са обозначени според стандартите.

Резултатите и приносите на проведеното дисертационно изследване са оригинални и не са заимствани от изследвания и публикации, в които нямам участие.

Библиография

1. Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6), 734-749.
2. Agarwal, D. (2013). Recommending Items to Users: An Explore Exploit Perspective. *The 1st workshop on User engagement optimization*.
3. Ahn, J., Brusilovsky, P., Grady, J., He, D., & Syn, S. (2007). Open User Profiles for Adaptive News Systems: Help or Harm? *Proceedings of the 16th International Conference on World Wide Web*, 11-20.
4. Anil, R., Dunning, T., & Friedman, E. (2011). *Mahout in action*.
5. Antulov-Fantulin, N., Bošnjak, M., Znidaršic, M., Grcar, M., Morzy, M., & Šmuc, T. (2011). ECML-PKDD 2011 Discovery Challenge Overview.
6. Balabanovic, M., & Shoham, Y. (1997). FAB: Content-based, Collaborative Recommendation. *Communications of the Association for Computing Machinery*, 66-72.
7. Bell, R. M., & Koren, Y. (2007). Lessons from the Netflix prize challenge. *ACM SIGKDD Explorations Newsletter*, 75-79.
8. Bennett, J. & (2007). The netflix prize. In Proceedings of KDD cup and workshop.
9. Billsus, D., & Pazzani, M. (2000). User modeling for adaptive news access. . *User modeling and user-adapted interaction*, 147-180.
10. Booth, P. (2014). Facebook reveals news feed experiment to control emotions. *The Guardian*.
11. Bostandjiev, S., O'Donovan, J., & Höllerer, T. (2012). TasteWeights: A Visual Interactive Hybrid Recommender System. *ACM RecSys'12*.
12. Breese, J. S., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*.
13. Burke, R. (2000). Knowledge-based recommender systems. *Encyclopedia of library and information systems*, 69(Supplement 32), 175-186.
14. Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 331-370.

15. Burke, R. (2007). Hybrid Web Recommender Systems. От *The adaptive web*. Springer.
16. Chen, W., & Fong, S. (2011). Social Network Collaborative Filtering Framework and Online Trust Factors: a Case Study on Facebook. *International Journal of Web Applications, Volume 3*.
17. Chen, W., & Fong, S. (November 2011). Social Network Collaborative Filtering Framework and Online Trust Factors: a Case Study on Facebook. *International Journal of Web Applications, Volume 3*.
18. Deshpande, M., & Karypis, G. (2004). Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems*.
19. Dover, D. (2008). *Reddit, Stumbleupon, Del.icio.us and Hacker News Algorithms Exposed!* Свалено от <http://moz.com/blog/reddit-stumbleupon-delicious-and-hacker-news-algorithms-exposed>
20. Dror, G. K. (2012). The Yahoo! Music Dataset and KDD-Cup'11.
21. Dunbar, R. (2010). *How many friends does one person need?: Dunbar's number and other evolutionary quirks*. Faber & Faber.
22. Ekstrand, M., Ludwig, M., Kolb, J., & Riedl, J. T. (2011). Lenskit: a modular recommender framework. In *Proceedings of the fifth ACM conference on Recommender systems*.
23. *Facebook 1 Billion stats*. (06.2013). <http://newsroom.fb.com/download-media/4227>
24. *Facebook Statistics*. (01.2014). <http://www.statisticbrain.com/facebook-statistics/>
25. Felfernig, A., Friedrich, G., Jannach, D., & Zanker, M. (2006). An integrated environment for the development of knowledge-based recommender applications. *International Journal of Electronic Commerce*, 11-34.
26. Funk, S. (2006). *Netflix update: Try this at home*. (06.2014) <http://sifter.org/~simon/journal/20061211.html>
27. Ge, H. (09 2013 r.). *News Feed FYI: More Relevant Ads in News Feed*. <http://newsroom.fb.com/news/2013/09/news-feed-fyi-more-relevant-ads-in-news-feed/>

28. Golbeck, J., & Hendler, J. (2006). Filmtrust: Movie recommendations using trust in web-based social networks. *Proceedings of the IEEE Consumer communications and networking conference*.
29. Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12), 61-70.
30. Herlocker, J., Konstan, J. A., & Riedl, J. (2002). An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms. *Information retrieval*, 5(4), 287-310.
31. Herlocker, J., Konstan, J., & Riedl, J. (2000). Explaining collaborative filtering recommendations. *ACM conference on Computer supported cooperative work*, (стр. 241-250).
32. Herlocker, J., Konstan, J., Terveen, L., & Riedl, J. (2004). Evaluating collaborative filtering recommender systems.
33. Ittner, D., Lewis, D., & Ahn, D. (1995). Text Categorization of Low Quality Images. *Symposium on Document Analysis and Information Retrieval*, 301-315.
34. Kazienko, P., & Musiał, K. (2006). Recommendation Framework for Online Social Networks. *Advances in Web Intelligence and Data Mining*, pp.111-120.
35. Koren, Y. (2008). Factorization meets the neighborhood: a multifaceted collaborative filtering model. *In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*.
36. Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 30-37.
37. Koychev. (2000). Gradual forgetting for adaptation to concept drift. *ECAI 2000 Workshop on Current Issues in Spatio-Temporal Reasoning*.
38. Koychev, I., & Schwab, I. (2000). Adaptation to drifting user's interests. *ECML2000 Workshop: Machine Learning in New Information Age*, (стр. 39-46).
39. Krulwich, B. (1997). Lifestyle Finder: Intelligent User Profiling Using Large-Scale Demographic. *AI Magazine* 18 (2), 37-45.
40. Lemire, D., & Maclachlan, A. (2007). Slope One Predictors for Online Rating-Based Collaborative Filtering.

41. Lops, P., de Gemmis, M., & Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. От *Recommender Systems Handbook* (стр. 73-105). Springer US.
42. M., P., & Billsus, D. (1997). Learning and Revising User Profiles: The Identification of Interesting. *Machine Learning* 27(3), 313-331.
43. Massa, P., & Avesani, P. (2007). Trust-aware recommender systems.
44. McCallum, A., & Nigam, K. (1998). A comparison of event models for naive bayes text classification. *AAAI-98 workshop on learning for text categorization* , 41-48.
45. Niu, Y., Wang, Y., Sun, G., Yue, A., Dalessandro, B., Perlich, C., & Hamner, B. (2012). The tencent dataset and kdd-cup'12.
46. Noel, J., Sanner, S., Tran, K. N., Christen, P., Xie, L., & Bonilla, E. V. (2012). New objective functions for social collaborative filtering. *Proceedings of the 21st international conference on World Wide Web*, (стр. 859-868).
47. O'Reilly, T. (2007). What is Web 2.0: Design patterns and business models for the next generation of software. *Communications and Strategies* 65.1, 17-37.
48. Pazzani, M. (1999). A hybrid user model for news story classification. *User Modeling*, 99-108.
49. Pazzani, M. J. (1999). A Framework for Collaborative, Content-Based and Demographic Filtering. *AI Review* 13 (5/6), 393-408.
50. Pazzani, M. J., & Billsus, D. (2007). Content-based recommendation systems. От *The adaptive web* (стр. 325-341). Springer.
51. Pazzani, M., & Billsus, D. (1997). Learning and Revising User Profiles: The Identification of Interesting. *Machine Learning* 27(3), 313-331.
52. Phillips, S. (25 07 2007 r.). A brief history of Facebook. The Guardian.
53. Pu, P., & Chen, L. (2006). Trust building with explanation interfaces. *The 11th international conference on Intelligent user interfaces*, (стр. 93–100).
54. Quijano-Sanchez, L., Recio-Garcia, J. A., & Diaz-Agudo, B. (2011). HappyMovie: A Facebook Application for Recommending Movies to Groups. *23rd IEEE International Conference on Tools with Artificial Intelligence*, (стр. 239-244).
55. Resnick, P., & Varian, H. (1997). Recommender systems. *Communications of the ACM*.

56. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). GroupLens: An Open Architecture for Collaborative Filtering of Netnews. *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*.
57. Rocchio, J. (1971). Relevance Feedback Information Retrieval. От *The SMART retrieval system - experiments in automated document processing*. Prentice-Hall.
58. Rushe, D. (2014). Facebook posts record quarterly results and reports \$1.5bn profit for 2013. *The Guardian*.
59. Salihefendic, A. (2010). *How Hacker News ranking algorithm works*. Свалено от <http://amix.dk/blog/post/19574>
60. Salihefendic, A. (2010b). *How Reddit ranking algorithms work*. Свалено от <http://amix.dk/blog/post/19588>
61. Salton, G., Wong, A., & Yang, C. (1975). A vector space model for automatic indexing. *Communications of the ACM*.
62. Schafer, J. B., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems. От *The adaptive web* 291-324
63. Schafer, J., Konstan, J., & Riedl, J. (1999). Recommender systems in e-commerce. *Proceedings of the 1st ACM conference on Electronic commerce*.
64. Servia-Rodríguez, S., Díaz-Redondo, R. P., Fernández-Vilas, A., & Pazos-Arias, J. J. (2013). Mining Facebook Activity to Discover Social Ties: Towards a Social-Sensitive Ecosystem. *In Cloud Computing and Services Science* 71-85.
65. Shani, G., & Gunawardana, A. (2011). Evaluating recommendation systems. От *Recommender systems handbook*. Springer.
66. Sinha, R. R., & Swearingen, K. (2001). Comparing Recommendations Made by Online Systems and Friends. *In DELOS workshop: personalisation and recommender systems in digital libraries*.
67. Sreedharan, A. (11 2013 r.). *Find your Facebook Friends Ranking Score !* Свалено от <http://arjunsreedharan.org/post/65979958297/find-your-facebook-friends-ranking-score>
68. Takács, G., Pilászy, I., Németh, B., & Tikk, D. (2009). Scalable collaborative filtering approaches for large recommender systems. *The Journal of Machine Learning Research*.
69. *UN's Millenium Development Goals Report*. (2013).United Nations.

70. Veritasium. (02.2014). *Facebook Fraud* . Свалено от <https://www.youtube.com/watch?v=oVfHeWTKjag&list=PL16649CCE7EFA8B2F>
71. Zanker, M., Felfernig, A., & Friedrich, G. (2011). *Recommender systems: an introduction*. Cambridge University Press.
72. Zhou, Y., Wilkinson, D., Schreiber, R., & Pan, R. (2008). Large-scale parallel collaborative filtering for the netflix prize. In *Algorithmic Aspects in Information and Management*, 337-348.