2014

Софийски университет „Св. Климент Охридски”
Факултет по математика и информатика
Катедра ”Информационни технологии”

Милен Митков Чечев

Система за препоръчване на съдържание в социалната мрежа

АВТОРЕФЕРАТ

на

дисертационен труд

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

Дисертационният труд е обсъден и насочен за защита на разширено заседание на катедра "Информационни технологии" към Факултета по математика и информатика на Софийски университет "Св. Климент Охридски", състояло се на 13.06.2014 г.

Авторът е докторант в задочна форма на обучение, като за същият период работи и като асистент във Факултета по математика и информатика на Софийски университет "Св. Климент Охридски".

Дисертационният труд е изложен на 91 страници, в които се съдържат 26 фигури, 4 таблици, 6 страници литература, включващи 72 заглавия. Списъкът от публикации на автора по същността на дисертацията включва 8 заглавия.

Публичната защита на дисертационния труд ще се състои на открито заседание на …….. г. от …….. часа в …….. Материалите по защитата са на разположение в …….. на ФМИ към СУ (София, бул. Джеймс Баучър №5).

Съдържание

Индекс на фигурите

Индекс на таблиците

Обща характеристика на дисертационния труд

Актуалност на проблема и мотивация

Препоръчващите системи започват развитието си в средата на 90-те години на миналият век. Интересът към тях е доста голям заради комерсиалната им насоченост и голямата възвращаемост на инвестицията. Интернет магазините, които предлагат препоръки на потребителите си, успяват да покажат на потребителя по широк асортимент от продукти, които биха му били полезни, и да увеличат приходите си. Пример за важността на препоръчващите системи за електронната търговия е състезанието „Наградата на Нетфликс", в което компанията за продаване на филми Нетфликс предлага 1 милион долара за отбора, който успее да подобри с 10% препоръчващата им система.

След първоначалния бум на препоръчващите системи за целите на електронната търговия в последните години се забелязва преориентиране на изследователската общност към препоръки в социалните мрежи. Това преориентиране се дължи най-вече на непрекъснато нарастващата популярност на социалните мрежи, като почти всички потребители на интернет имат регистрация в някоя социална мрежа.

Според статистиките в края на 2013, интернет потребителите са около 2,7 милиарда[*], а в социалната мрежа Фейсбук активните потребители вече са повече от 1.3 милиард като само за последната една година са се увеличили с 22%[†].

Социалната мрежа Фейсбук е създадена през 2004 година като достъпна и лесна среда за комуникация на студентите в Харвард. В последствие възможността да се споделя информация и да се поддържат социални контакти я разпространява из целият свят и днес тя се е превърнала в леснодостъпен форум, където милиони хора от цял свят ежедневно споделят снимки, лична информация, мисли и мнения по различни социални теми. Достъпността на споделянето на информация, обаче е както привлекателно, така и създава проблем – потребителите получават прекалено много информация, което затруднява разглеждането на целият и обем. Според статистиката средният брой на приятелите на потребителите регистрирали се във Facebook от повече от 2

[*] *UN's Millenium Development Goals Report.* United Nations. 2013
[†] http://www.statisticbrain.com/facebook-statistics/

години е 305[‡]. От друга страна Робин Дънбар прави изследване и показва, че човешкият мозък има физическо ограничение за броя на социалните връзки, които може да поддържа. Той определя 150 като максимален брой на връзките, които човек може да поддържа. Наличието на много социални връзки и лесният интерфейс за споделяне на информация довежда до създаването на прекалено голям поток от социална информация като потребителят не може да отсее какво наистина го интересува и какво може да пропусне. Подобен е основният проблем на извличането на информация (information retrieval), заради който Google придобива популярност в края на 20 век. Приликата е, че потребителят се нуждае бързо и лесно да се сдобие с релевантна информация, като в търсещите системи задачата е да се сортират резултати от една заявка, докато в социалната мрежа задачата е да се сортира потока от социални новини.

Обект на изследването е ежедневното взаимодействие на потребителите в социалната мрежа с техните приятели и с информацията споделена в социалната мрежа. Изследва се възможността да се подреди социалният 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 му алтернативен изглед на неговият социален поток.

Изследователският въпрос, който е поставен в настоящият труд е: „Как да се подреди социалният поток на 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.

[‡] *Facebook 1 Billion stats.* http://newsroom.fb.com/download-media/4227

Задачи на дисертационния труд

Задачите на дисертационния труд са следните:

- Обзор на съществуващите подходи за препоръчване на информация в социалните мрежи
- Изследване на необходимостта от нова препоръчваща система, за препоръчване на информация в социалната мрежа.
- Проектиране на подход за препоръчване на информация базиран на резултатите от предишната точка.
- Оценяване на ефективността на предложеният подход.

Структура и съдържание на дисертационния труд

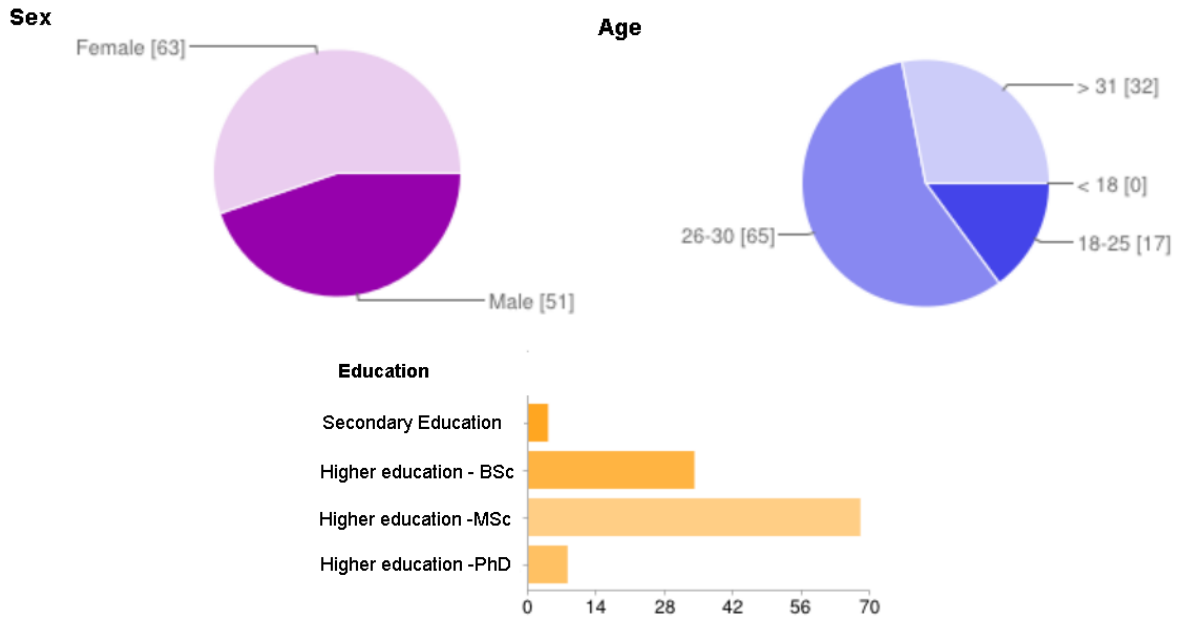
Дисертационният труд е с общ обем XXX страници. Състои се от увод, изложение в XX глави, заключение, списък с цитираната литература . Дисертационният труд е описан на XXX страници, в които се съдържат XXX фигури, XXX таблици, XXX страници литература, включващи XXX заглавия, от които XXX на български език, XXX на английски език и XXX интернет източника. Списъкът от публикации на автора по същността на дисертацията включва XXX заглавия.

Глава 1. Обзор

В **първа глава** е направен обзор на проблемната област. Разгледани са различните видове препоръчващи подходи: препоръки базирани на съдържание, препоръки базирани на сътрудничество, препоръки базирани на демографски фактори, препоръки базирани на знания, препоръки базирани на общност и хибридни препоръчващи системи. Направен е обзор на методите за оценяване на препоръчващите системи и техните свойства. В секцията софтуерни инструменти и платформи са споменати двете най-известни платформи за препоръчващи системи Apache Mahout и Lens Kit и са описани програмните интерфейси на Facebook и Twitter, като две от най-големите социални мрежи в Европа и Северна Америка.

Глава 2. Изследване и анализ на проблема

Във **втора глава** е направено изследване за необходимостта на потребителите на социалната мрежа Фейсбук от система за препоръчване на информация. Направено е допитване до 114 активни потребители на социалната мрежа със следните демографски данни



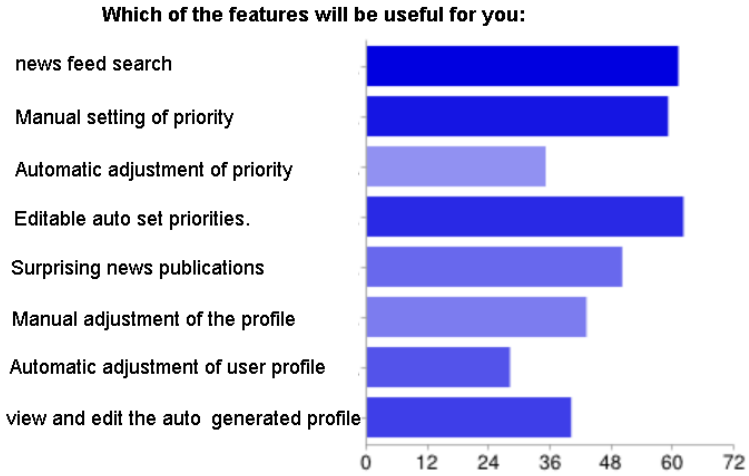
Фигура 1. Демографски харектеристики на участниците в анкетата

Попитани за времето, което прекарват във фейсбук 25% отговарят, че не са много редовни потребители и влизат от време на време за по 5 мин, но останалите 75% влизат всеки ден като 47% отговарят, че отделят около половин час на ден, а 25% по-повече от час на ден.

Потребителите определят основната си активност във Фейсбук като разглеждане на потока с новини, като 62% от хората определят, че тя заема повече от половината от времето, което прекарват 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 2. Избор на нова функционалност от Фейсбук потребителите.

Потребителите показват доста слаба активност на 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.

Изводите направени като част от втора глава са:

- Има необходимост от приложение подпомагащо потребителите в избора им на новини от социалната мрежа.
- потребителите са скептично настроени към препоръчващите системи за това е необходимо приложението да обяснява резултатите си и да има опции за ръчна настройка.
- За да е по-атрактивно, приложението трябва да предлага и опции за търсене на информация

Грава 3. Извличане на базата от знания от експертите

В трета глава са разгледани основните препоръчващи подходи изследвани в дисертацията. Изследвани са подходи с препоръчване базирано на съдържание, сътрудничество, доверие, както и хибриден подход. Допълнително е разгледан и проблема за остаряване на новините с времето.

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, 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 Рочио, 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 от неговото съдържание и създаденият потребителски профил, като за метрика за сравнение на двата вектора използваме формулата за косинусова прилика:

$$similarity(d_a, u) = \frac{d_a . u}{|d_a||u|} = \frac{\sum_{i=1}^{m} d_{a,i}\, u_i}{\sqrt{\sum_{i=1}^{m} d_{a,i}^2}\,\sqrt{\sum_{i=1}^{m} u_i^2}}$$

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.

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 има достъп до 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 при изготвяне на препоръки за потребител се използват само данните събрани от приложението с правата дадени му от потребителя. Това са публикуваните от него и неговите приятели 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.

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

$$similarity(user_k, user_l) = \frac{\sum_{i=1}^{N} M_{ki} M_{li}}{\sqrt{\sum_{i=1}^{N} M_{ki}} \sqrt{\sum_{i=1}^{N} M_{li}}}$$

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

Използвайки тази формула за близост, най-близки до нашият потребител ще бъдат потребители, които са харесали много обекти заедно с нашият потребител 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 Пиърсън, 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 нас, то те не бива да са близки до нас. Формулата, която използваме прави тази нормализация разделяйки на броя на харесаните от потребител обекти.

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.

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

Препоръки базирани на доверие

В мрежите базирани на доверие като Epinions и FimTrust средният брой връзки на доверие е около 1.7[§]. 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, като гъстотата на мрежата е доста по-голяма. Средно потребителите в мрежата имат около 300 приятели, което налага връзките на приятелството да бъдат оценени и сортирани спрямо степента на доверие на потребителя в публикациите на приятелите му.

За да се изчисли степента на доверие се използва информация за активността на потребителя и взаимодействието с приятелите му. Чен и Фонг[**] дефинират доверието като симетрична променлива, но в реалния свят доверието често е асиметрично за това в дисертацията го дефинираме като асиметрична 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 се прави изчислението за доверие е следната:

$$trust_{ij} = \frac{\alpha.like_{ij} + \beta.comment_{ij} + \gamma.share_{ij}}{published_j} + \mu.\frac{activity_{ij}}{activity_i}$$

[§] Massa, P., & Avesani, P. (2007). Trust-aware recommender systems. In Proceedings of the 2007 ACM conference on Recommender systems (pp. 17-24). ACM.

[**] 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.

Във формулата с trust_{ij} отбелязваме доверието на потребител i в потребител j, $like_{ij}$ е броят 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. Моделът се обучава така, че параметрите α, β, γ, μ 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:

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

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 $\max(\text{trust}_i)$, което е максималното доверие което може да получи един обект. Максималното доверие е равно на акумулираното доверие от всички приятели на потребителя:

$$max(trust_i) = \sum_{j \in friends_i} trust_{ij}$$

Отчитане на времевият 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 News[††], Reddit[‡‡] and Hacker news[§§]. 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 Hacher 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:

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

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 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 някой видове. Във формулата на Hacker news се използва и променливата t_{now}, която е времевият отпечатък на текущият момент в който се изчислява оценката. Формулата, която използват е следната:

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

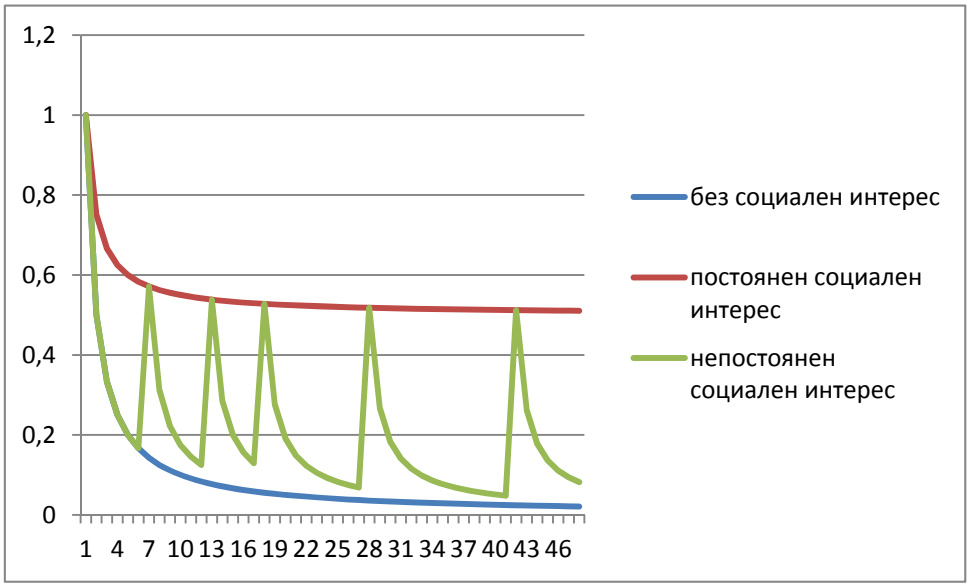
За направа на оценка на различните новини в социалната мрежа, първо се използват алгоритмите описани в предишните точки и след това се прилага стратегия за забравяне по-подобен начин на забравянето внедрено в представените два новинарски сайта. Важно е да се отбележи, че в социалната мрежа, често новина е интересна заради добавен нов коментар или отбелязано харесване към нея. Поради тази причина, при остаряване на новината, трябва да вземем в предвид едновременно кога е публикувана и кога последно някой от

[††] http://news.google.com/
[‡‡] http://reddit.com/
[§§] http://news.ycombinator.com/

приятелите ни е взаимодействал с нея. Допълнително ако една новина е публикувана и в последствие никой 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, 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 времевият отпечатък в момента.

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

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 3 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.



Фигура 3. Времева промяна на оценката с различна степен на социален интерес.

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.

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 тази формула:

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

Глава 4: Система за препоръчване в социалната мрежа

В четвъртата глава е разгледана в детайли архитектурата, функционалността и потребителският интерфейс на приложението. Описана е следната функционалност, необходима за подобно приложение:

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

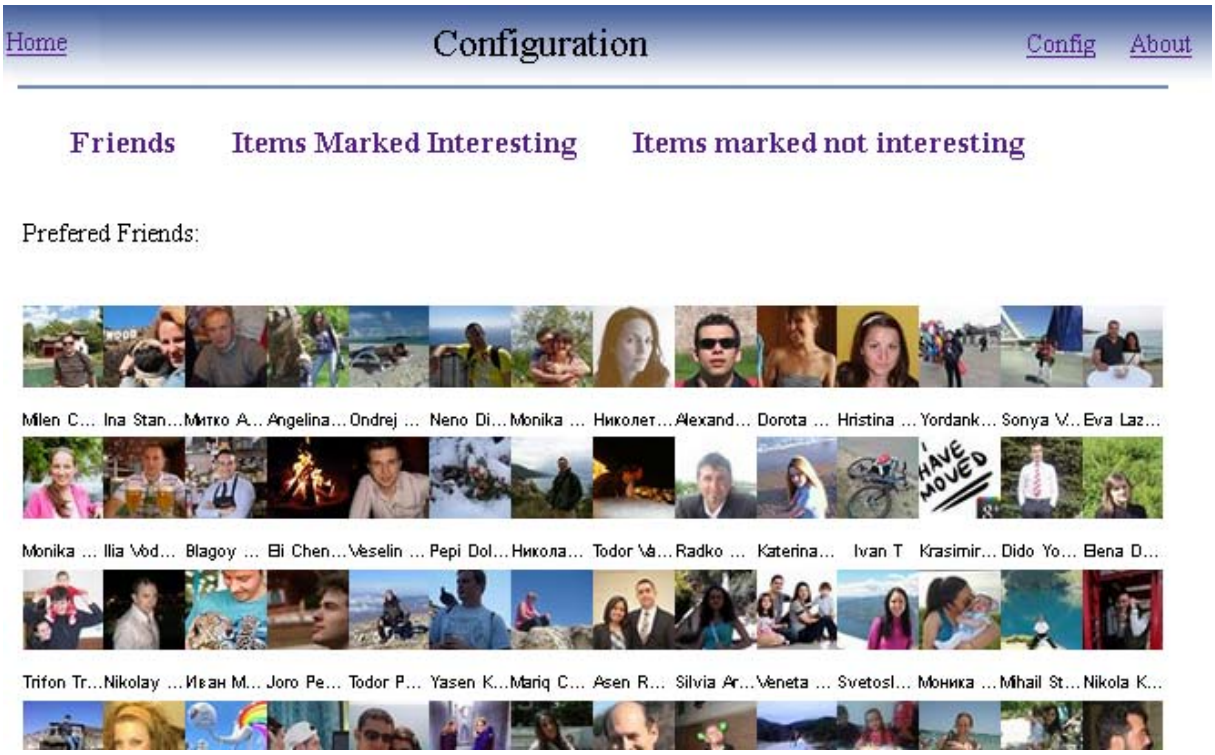
Използваната архитектура (Фигура 4) използва следните основни модули: модул за връзка и комуникация със социалната мрежа, модул за обработка на извлечената информация и съхраняването и в хранилище за данни, модул за генериране на потребителски профил от съхранената информация, модул за генериране на списъка с препоръки, модул за комуникация с потребителя(потребителски интерфейс), модул за съхранение на информацията предоставена от потребителя като обратна връзка.

Фигура 4. Архитектура на системата.

Потребителски интерфейс на приложението предлага цялата, функционалност необходима за четене на новини и работа с потока от информация - Фигура 5.

Фигура 5. Основен изглед от приложението.

Допълнително приложението предлага на потребителите възможност за разглеждане на създаденият им профил на приятелствата и за финна ръчна настройка - Фигура 6



Фигура 6. Потребителски профил.

Глава 5: Оценяване

В пета глава е представено направеното оценяване на точността на направените препоръки и удовлетвореността на потребителите от нея. 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 с неявна и явна обратна връзка.

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 32 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 приятеля 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 и се 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 1

Брой регистрирани потребители	32
Брой на приятелите	9462
Публикации от поток с новини	10710
Публикации публикувани от приятели	853366
Харесани обекти	19212

Таблица 1. Статистики за събраните данни.

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 или 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.

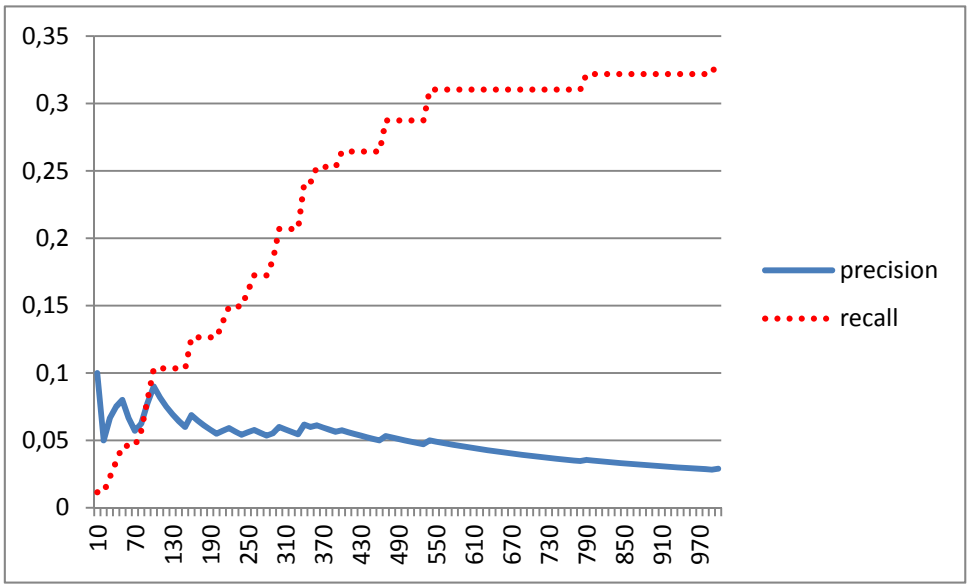
За проведените експеримент е използвана 10 групова (10-fold) кросвалидация. Като сравняваните подходите са препоръки базирани на съдържание, препоръки базирани на сътрудничество, препоръки базирани на доверие и хибриден подход използващ паралелна архитектура обединяващ трите алгоритъма. При комбиниране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

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 изчисление се използва средно аритметичната стойност от всички резултати:

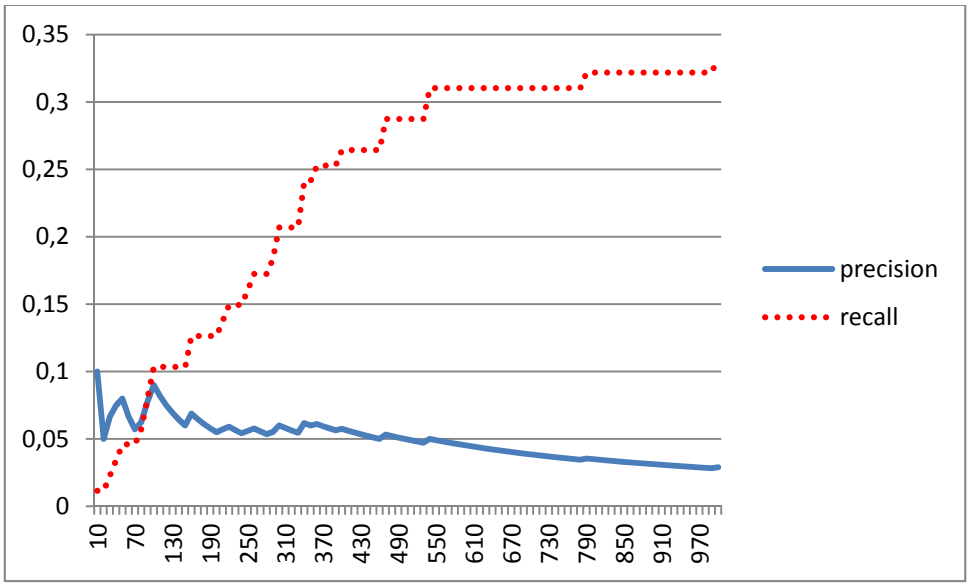
$$mean\_precision@N = \frac{\sum_{for\ all\ users} precision@N}{\#number\ of\ users}$$

$$mean\_recall@N = \frac{\sum_{for\ all\ users} recall@N}{\#number\ of\ users}$$



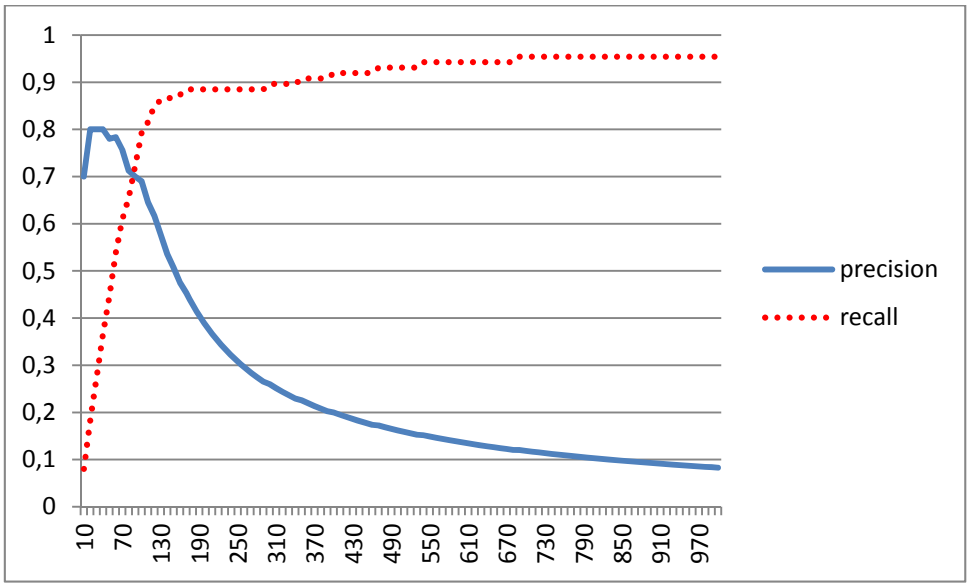
Фигура 7. Точност и откриване с препоръки базирани на съдържание.

На Фигура 7 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 8. Точност и откриване с препоръки базирани на съдържание.

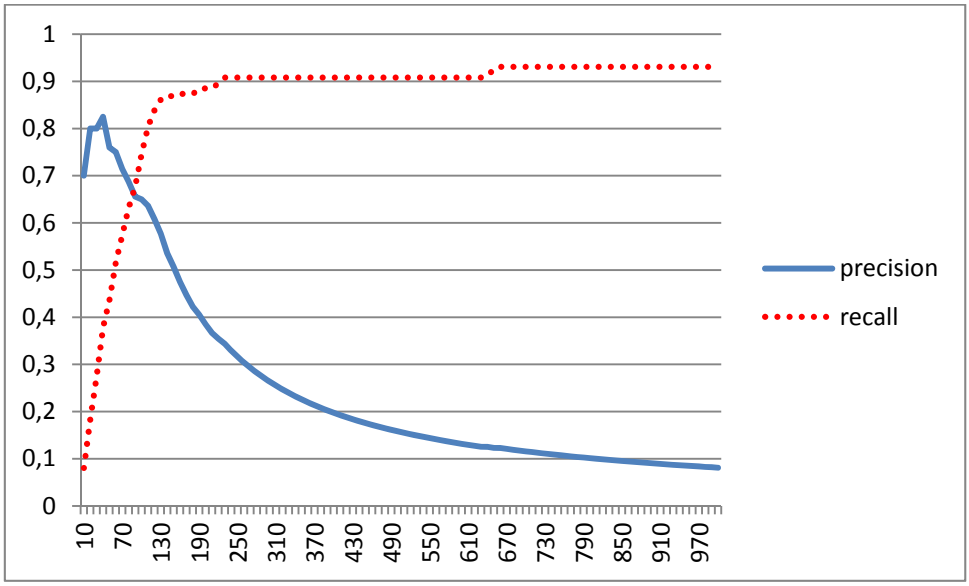
Резултатите от препоръките базирани на сътрудничество са показани на Фигура 9. Забелязва се, че точността е значително по-висока от тази на препоръките базирани на съдържание като при препоръчване на 20 обекта имаме точност от 0.8, а при препоръчване на 100 обекта имаме откриване 0.8.



Фигура 9. 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 някой разлики с препоръките, базирани на сътрудничество. При препоръчани 40 обекта препоръките базирани на доверие имат по-висока точност от тази на

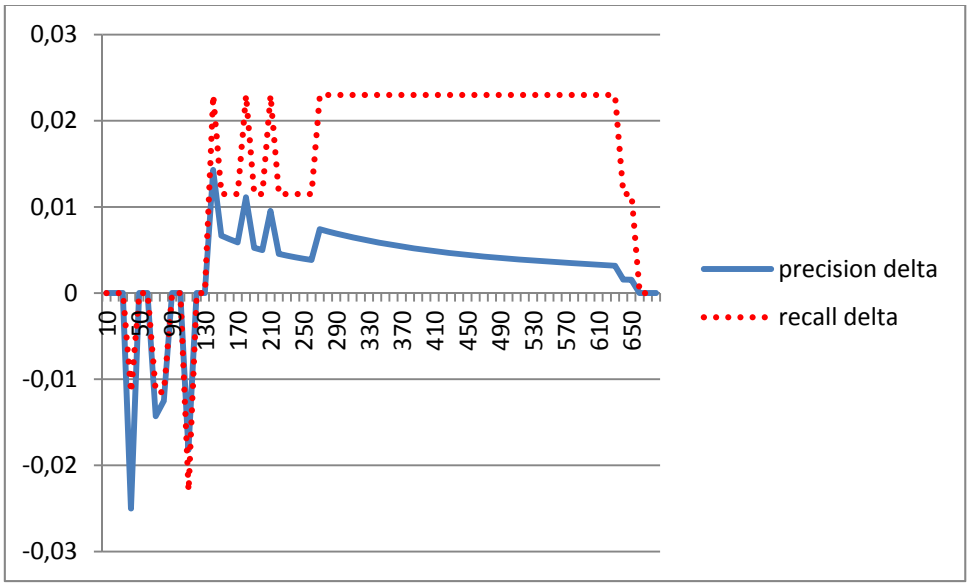
препоръките базирани на сътрудничество – 0.825 срещу 0.8. Препоръките базирани на сътрудничество показват обаче по-висока точност при разглеждане на първите 70 препоръки – 0.757 срещу 0.714. Това поведение на по-силни препоръки в началото на списъка и не чак толкова силни следващи резултати е очакван 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 10. Точност и откриване на препоръки базирани на доверие.

При изследването на хибридния подход не са наблюдавани големи разлики с резултатите от подхода базиран на доверие, поради което за по-добра визуализация на Фигура 11 представяме разликата между метриките на подхода базиран на доверие и хибридният подход. Вероятно липсата на подобрение в резултатите на хибридният алгоритъма 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 впечатления, 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 11. 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.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:

$$mean\_nDCG = \frac{\sum_{for\ all\ users} nDCG}{\#number\ of\ users}$$

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

$$MAP = \frac{\sum_{for\ all\ users} AveP(u)}{\#number\ of\ users}$$

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

Таблица 2 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 2. 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.

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 3 представя въпросите и дадените отговори в проценти. Където за въпросите е използвана ликертова скала с възможни отговори: 1=Твърдо не; 2=По-скоро Не; 3=Колебая се; 4=По-скоро да; 5=Да.

Въпроси:	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 в 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 променята наредбата на приятелите си в потребителския профил?	0.0%	0.0%	0.0%	25.0%	75.0%
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 3. 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 потребителският 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.

Заключение

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 вицовете, но не и новините, докато от друг може да е точно обратното.
- Различени 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.

Авторска справка

Основните приноси на дисертационния труд могат да бъдат разделени на две групи - научни и научно-приложни.

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, AIMSA'2014, September 2014 (приета 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.

.

Кратка автобиография

Милен Чечев работи като щатен преподавател-асистент във ФМИ, СУ през периода септември 2007-юни 2014 през това време той участва в екипите на курсовете „Изкуствен интелект”,”Извличане на информация”, „Извличане на знания от данни”, „Структури от данни и програмиране”, „Обектно Ориентирано програмиране”,”Увод в програмирането” и създава курс свързан по тематиката на дисертацията си „Препоръчващи системи”, който се провежда през 2012 и 2014 с общо 50 обучавани студенти в него.

Милен започва работа по докторантурата си през юни 2009, като през периода септември 2010-март 2011, той специализира в Aalto University, Финландия с ръководител професор Erkki Oja. В последствие през периода март 2011-юли 2012 той работи допълнително в компанията Онтотекст, където ръководи изследователската и развойната дейност по проектите от 7-ма рамкова програма MOLTO и INSEMTIVES. През периода февруари 2013-февруари 2014 работи в Амазон Канада, където извършва изследователска дейност с анализ на големи масиви с данни за целите на оптимизация на процесите в компанията.

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 ръководил в това направление.