ИЗПОЛЗВАНЕ НА СТРУКТУРИРАНА ИНФОРМАЦИЯ ОТ ТАГОВЕ ЗА ПРЕПОРЪЧВАНЕ НА КНИГИ

Разширено резюме на български език

От Мелания Бербатова, магистърска програма “Изкуствен интелект”
Факултет по математика и информатика,
Софийски университет “Св. Климент Охридски”

Февруари 2020

Резюме

Препоръчващите системи са софтуерни системи, базирани най-често на алгоритми за машинно самообучение, които се използват за препоръчване на продукти и онлайн съдържание на онлайн потребители, като препоръките се базират на научаването на потребителския профил и преференциите на потребителите. Препоръчващите системи навлизат все по-често в нашия живот, като до голяма степен са движеща сила на големите сайтове за продажби и разпространение на онлайн съдържание, като *Amazon*, *Youtube* и *Netflix*.
В настоящата дипломна работа разглеждаме препоръчващи системи за книги, като базираме работата си на свободно наличните данни *goodbooks-10k* от сайта *GoodReads.com*, от които разглеждаме най-вече потребителските тагове. Потребителските тагове отразяват най-разнообразна неструктурирана инфомация, като мнение и отношение на потребителя към книгата, и характеристики на книгата като поредица, автор, жанр и други. Целта на работата ни е да извлечем структурирана информация от таговете под формата на граф, която чрез алгоритми за графови ембединги да приведем във векторно представяне и да използваме за създаването на препоръчваща система, базирана на съдържание. За тази цел провеждаме експерименти с няколко подхода за научаване на структура от таговете - научаване на връзки между таговете чрез ембединги на думи, клъстеризиране и съпоставяне с експертна онтология. Екпериментите показват подобрение на препоръките при използването на графова структура спрямо използването на данните в суров формат. Резултатите ни дават надежда, че разработването и разширяването на методи, подобни на изследваните, може да доведе до по-добри препоръки и да бъде използвано в препоръчващи системи, използвани за реални цели.

Цели и задачи на дипломната работа

Основната цел на тази дипломна работа е да разработи система, която може ефективно да използва потребителските тагове на книги за даване на подходящи препоръки. За постигане на тази цел дефинираме набор от задачи, които трябва да бъдат изпълнени:

1. *Преглед на областта*. Тази задача включва търсене на научни статии, статии от блогове, проекти с отворен код и всякаква друга работа, извършена по темата за използването на структуриране информация от тагове и за препоръчване на продукти, и по-специално на книги. Резултатите, които откритите изследвания постигат, трябва да се анализират прецизно, за да могат да бъдат от полза за развитие на нашата система.

2. *Проучване на данните.* Тъй като наборът от данни е разнороден и обемен, трябва да внимателно да разгледаме характеристиките на динните и разработим модели за създаване на по-малък набор от данни за експерименти. Тази задача включва и търсене на външни ресурси, които са подходящи за нашата задача, например бази от знания и онтологии, например.

3. *Избор на набор от подходи.* Тъй като задачата на препоръките за книги е много широка, ние трябва да изберем набор от подходи, с които ще работим. Въз основа на нашите изследвания, можем да определим списък от задачи и да извършим постепенни експерименти.

4. *Изграждане на система за препоръки.* Тази задача включва избор на архитектурата на препоръката, избиране на алгоритъм за учене и избор на метрики за оценка.

5. *Конструиране на признаци (feature engineering)*. В тази задача трябва да организираме информацията за книгите във форма, подходяща за архитектурата на препоръчващата система. За тази цел ще създадем
различни набори от признаци, които ще бъдат подадени към системата.

6. *Планиране и изпълнение на експериментите.* В тази задача ще определим експериментите, които ще изпълним. Ще изберем един модел за основен и ще опитаме постепенно да го подобрим.

7. *Оценка и интерпретация на резултатите.* В края на работата си трябва да анализираме и сравним разработените модели, както и да интерпретираме резултатите и да обясним защо някои подходи дават по-добри резултати от други.

Структура и съдържание

Настоящата дипломна работа е разделена в следните 6 глави:

1. Въведение.

Във Въведението е разгледана накратко проблемната област - препоръчващи системи за книги и е посочена мотивация за развитието на изследвания в областта. Посочени и илюстрирани са няколко примера за препоръки на книги от сайта *Goodreads.com*. Накрая са дефинирани целите и задачите на дипломната работа, които посочихме по-горе.

Фиг. 1. Примери за препоръки на сайта Goodreads.com. Забелязваме, че настоящите алгоритми не се справят добре с препоръчването на подходящи книги. На първата препоръка книгата е на език, различен от този, на който потребителят чете, а на втората - препоръките са неперсонализирани.

2. Преглед на областта

Това е най-обширната глава от дипломната работа, която прави преглед на всички тематики, концепции и алгоритми, разгледани в текста. Състои се от следните части:

2.1. *Препоръчващи системи* - Разглеждат се същността на препоръчващите системи, основните видове - базирани на съдържание и колаборативни, както и най-често използваните алгоритми за реализирането им.

2.2. *Препоръчващи системи за книги* - Разглеждат се изследванията от последните години, свързани с препоръчване на книги, и по-специално тези, свързани с данните, използвани в дипломната работа. Направен е анализ на методите, използвани досега, и са предложни подобрения.

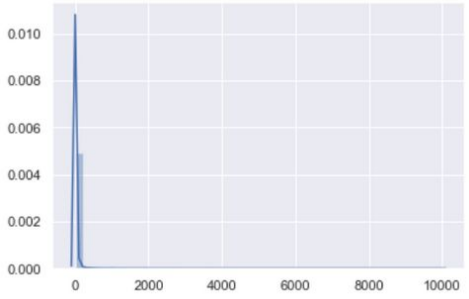
2.3 *Научаване на структурирано знание от тагове* - В тази секция се разглеждат основните методи за извличане на структурирана информация от тагове, изброяват се някои проблеми на тези методи и се предлага система за оценяването им. В края на секцията са разгледани по-специално характеристиките на таговете за книги.

2.4. *Ембединги на графи* - разглежда се проблема за научаването на представяния (*representation learning*) и по-специално на ембединги. Дефинира се проблемът за ембедингите на графи, който ние използваме в частта си за ембединги на възли на графа (*node embeddings*).

2.5. *Node2vec* - В последната секция от тази глава представяме алгоритъмът за ембединги на възли, с който сме избрали да работим - node2vec. Разглеждаме как той изгражда ембедингите, както и някои негови интересни характеристики.

3. Данни

В тази глава правим правим проучвателен анализ на данните, с които работим - goodbooks-10k, и извеждаме техни описателни статистики. Също така разглеждаме и външните данни, които използваме - онтология на жанрове.




```
count    34252.000000
mean        29.192806
std        277.254715
min          1.000000
25%          1.000000
50%          1.000000
75%          5.000000
max       9983.000000
Name: tag_name, dtype: float64
```

Фиг. 2. Разпределение на използването на таговете в набора от данни. Около половината от таговете са били използвани само за 1 книга, като най-често използвания таг ("to-read") е бил използван за почти всички книги. Тази статистика ни показва необходимост от предварителна селекция на данните

	genre	label
1	genre:realist	"realist"@en
2	genre:fairytale	"fairytale"@en
3	genre:guidebook	"guidebook"@en
4	genre:treatise	"treatise"@en
5	genre:novella	"novella"@en
6	genre:polemic	"polemic"@en
7	genre:magicRealist	"magic realist"@en
8	genre:panegyric	"panegyric"@en
9	genre:slaveNarrative	"slave narrative"@en
10	genre:dramaticMonologue	"dramatic monologue"@en

Фиг. 3. Примери за жанрове в онтологията

4. Дизайн на системата

В тази глава представяме модула за препоръчващи системи TensorRec, чрез който правим прототип на препоръчваща система. Обсъждаме дизайна на препоръчващата система, каква форма на данните и каква обучителна функция използва. Също така обясняваме как избираме и трансформираме данните, преди

да ги подадем на препоръчващата система за обучение и тестване. Представяме метриката за оценка, която използваме - recall@10, и аргументираме избора си.

5. Експерименти

В тази глава представяме направените експерименти и постигнатите резултати. Проведените експерименти се състоят от следните групи:

5.1. *Създаване на подмножество за тестване* - В началото създаваме подмножество от данните, върху което дефинираме разнообразни релации, които могат да съществуват между таговете. Дефинираните релации са:

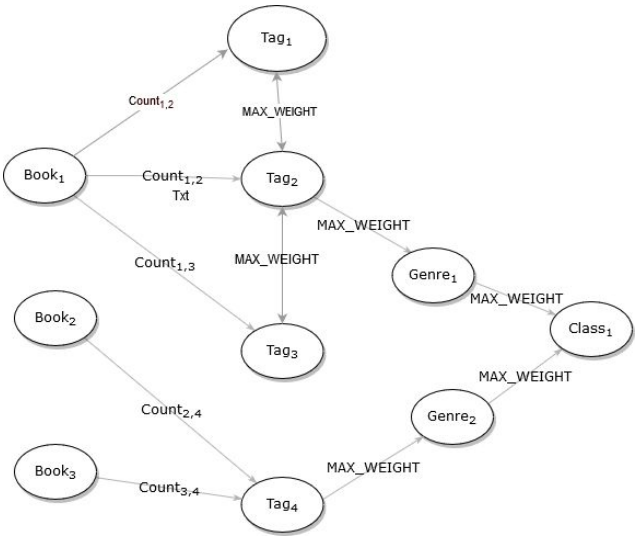
- синонимими,
- специфичност (хиперними/хипоними),
- съкращения,
- сродни думи,
- семантично разширение,
- форми на думата,
- различни изписвания едни и същи думи,
- еднакви наименовани единици.

5.2. *Ембединги на думи* - Разслеждаме особеностите на различни претренирани ембединги на думи - GloVe, ELMo, BERT - и това как се справят със задачата за научаване на връзки между таговете. Тестваме и сравняваме резултатите на тези методи върху тестовото множество, като накрая се спираме на комбиниран метод.

5.3. Клъстеризиране - След това разглеждаме и илюстрираме методи за клъстеризиране и откриване на йерархия между таговете - DBSCAN и съвместно появяване (*co-occurence*) на тагове. Коментираме проблемите на тези методи и на клъстеризането на тагове като цяло. Заради недоброто представяне, тези методи не са включени във крайните експерименти.

5.4. *Съпоставяне към външни ресурси* - Разглеждаме методи за построяване на нови връзки чрез съпоставянето на таговете към външните ресурси, с които работим: онтологията на жанрове и DBPedia. За съжаление, онтологията, която използваме има малко покритие на таговете, а скалирането на съпоставянето на таговете към концепти от DBPedia беше неуспешно поради проблеми в комуникацията със сървъра.

5.5 *Графови ембединги* - В този раздел обясняваме как използваме гореописаните методи за построяване на графова структура от данни. Графовата структура получаваме, като първоначално правим възли (*книга*, *таг*) от суровите данни. След това поетапно добавяме възли (*таг*, *таг*), (*таг*, *жанр*) и (*жанр*, *клас*). Представяме алгоритъм за тестване на графова структура, използващ графови ембединти и препоръчваща система.



Фиг. 4. Схематична илюстрация на възможни възли и ребра в получения граф

След това тренираме ембединги върху получените стуктури чрез алгоритъма *node2vec*. Представяме експерименти с графи с различни множества от възли и ребра и с ембединги с различни параметри. Накрая представяме таблично получените резултати и правим изводи. Виждаме, че най-добри резултати върху тестовото множество се получават при комбинирането на всички видове ребра, претеглен граф и използването на по-малък брой рандомизирани разходки.

Използвани възли	Брой рандомизирани разходки	Характеристики на графа		Recall@10	
		Претеглен	Насочен	Тренировъчни данни	Тестови данни
Сурови данни	-	-	-	0.0394	0.0193
Книга-таг	5	да	да	0.0412	0.0406
Книга-таг	100	да	да	0.0677	0.0389
Книга-таг	100	да	не	0.0565	0.0236
Книга-таг + таг-жанр	5	да	да	0.0463	0.0462
Книга-таг + таг-жанр + жанр-клас	5	да	да	0.0506	**0.0509**
Книга-таг + таг-жанр + жанр-клас	5	не	да	0.0113	0.0120
Книга-таг + таг-жанр + жанр-клас	100	да	не	0.0861	0.0356
Книга-таг + таг-таг	5	да	да	0.0623	0.0617
Книга-таг + таг-таг	100	да	не	0.0800	0.0424

Табл. 1. Резултати от проведените експерименти с графови ембединги

6. Изводи и бъдеща работа

В последната глава от дипломната работа прави изводи от проведените експерименти и даваме насоки за бъдеща работа с цел развитие на системата и подобряване на резултатите и.

Приноси

Дипломната работа съдържа следните приноси към развитието на областта:

1. Прави разширен преглед и анализ на изследваната област, като резюмира информация и изводи от различни изследвания.
2. Предлага и разглежда използването на методи, разработени през последните години, като ембединги на думи и графи, за задачите за научаване на структурирана информация от графи и оценяването и чрез препоръчваща система.
3. Дефинира възможни релации между тагове.
4. Предлага и описва цялостна рамка за създаване и оценка на препоръчваща система, базирана на тагове.

Резултати и изводи

Получените резултати показват, че структурирането на таговете в граф подобрява способността им да служат за препоръки, базирани на съдържание. Препоръчващата система, която предлагаме, показва по-добри резултати при използването на таговете в графови структури, спрямо тези при използването им в суров формат. Все пак, за да бъдат още по-точни и убедителни резултатите, е нужно те да бъдат тествани и с други алгоритми за препоръчващи системи, базирани на съдържание. Показаните методи дават насоки за развитието на бъдещи изследвания в областта, но все още не са достатъчно добри за директното им използване в задачи за препоръчващи системи от реалния свят.

Бъдещо развитие

В бъдеще, проектът може да бъде развит в следните насоки:

1. Дизайн на собствена препоръчваща система, вместо използваната за прототип, с цел по-голям контрол върху използваните алгоритми.
2. Използването на база от знания за структурирането на информация. Базата от знания ще даде допълнителна информация на графа - типове връзки и йерархия между възлите.
3. Разработване на хибридни методи за научаване на на структурирана информация от графа, тъй като никой от показаните методи не е достатъчно добър самостоятелно.
4. Използване на пълния набор от данни и тестването на ембедингите с повече алгоритми за препоръчващи системи, като регресия и k най-близки съседа.
5. Още настройки на параметрите, тъй резултатите от node2vec ембедингите зависят до голяма степен от настройките на параметрите му.