УДК 004.8

СИСТЕМЫ НАУЧНЫХ РЕКОМЕНДАЦИЙ

[1]Мамай И.Б., [2]Ильин Д.А., [2]Лимонова Е.Е., [3]Путинцев Д.Н.

[1]НИТУ «МИСиС», Москва;

[2]Московский физико-технический институт, Москва;

[3]Институт системного анализа ФИЦ ИУ РАН, Москва, e-mail: 2001dnp@mail.ru

Примерами научных рекомендательных систем являются реферативные системы, рекомендательные системы подбора рецензентов, цитатные рекомендательные системы и системы рекомендаций по научным статьям и журналам. Проблемы выработки рекомендаций применительно к научным объектам (например, научным статьям, журналам, научным сотрудникам) существенно отличаются от традиционных рекомендательных задач по потребительским товарам или фильмам. Предметом перспективных исследований в области рекомендующих систем являются модели, объединяющие предметные и семантические аспекты в единую инфраструктуру. Одним из эффективных методов выявления социальных связей между исследователями является кластер-анализ, основанный на сегментации тематических интересов. Статья знакомит с существующими подходами, применяемыми в системах научных рекомендаций, и дает описание современных алгоритмов формирования рекомендаций. Рассмотрены методы, базирующиеся на трех принципах: контент-ориентированная фильтрация, коллаборативная фильтрация и комбинированный метод.

Ключевые слова: научные рекомендации, контент-ориентированная фильтрация, коллаборативная фильтрация, гибридный подход

SYSTEMS OF SCIENTIFIC RECOMMENDATIONS

[1]Mamay I.B., [2]Ilyin D.A., [2]Limonova E.E., [3]Putintsev D.N.

[1]NUST MISiS, Moscow;

[2]Moscow Institute of Physics and Technology (State University), Moscow;

[3]Institute for Systems Analysis, FRC CSC RAS, Moscow, e-mail: 2001dnp@mail.ru

Examples of scientific advisory systems are referral systems, referee selection systems, citation advisory systems and recommendations systems for scientific articles and journals. The problems of making recommendations with respect to scientific objects (for example, scientific articles, journals, researchers) differ substantially from traditional recommendations on consumer goods or films. The subject of promising research in the field of recommending systems are models that integrate subject and semantic aspects into a single infrastructure. One of the effective methods for identifying social ties between researchers is a cluster analysis based on the segmentation of thematic interests. The article acquaints with the existing approaches used in the systems of scientific recommendations, and gives a description of modern algorithms for the formation of recommendations. Methods based on three principles are considered: content-oriented filtering, collaborative filtering, and a combined method.

Keywords: scientific advice, content-oriented filtering, collaborative filtering, hybrid approach

Главное предназначение систем рекомендаций заключается в поддержке навигации целевого пользователя по сложному информационному пространству. В основе выработки рекомендаций находится совокупность знаний системы о пользователе, других пользователях в системе, и самого информационного пространства. Все системы используют информацию о пользователе (иногда называемую профилем пользователя, пользовательской моделью или пользовательскими настройками) для формирования рекомендаций. Примерами научных рекомендательных систем являются реферативные системы, рекомендательные системы подбора рецензентов, цитатные рекомендательные системы и системы рекомендаций по научным статьям и журналам [12].

Проблема выработки рекомендаций применительно к научным объектам (например, научным статьям, журналам, научным сотрудникам) существенно отличается от традиционных рекомендательных задач по потребительским товарам или фильмам. Контент-ориентированные подходы, основанные на анализе содержимого, используют при определении соответствия предпочтениям потенциальных объектов, в основном, ключевые слова или фразы из предметной области, игнорируя при этом семантические связи (например, соавторство и цитирование).

Предметом перспективных исследований в области рекомендующих систем являются модели, объединяющие предметные и семантические аспекты в единую инфраструктуру [3]. Одним из эффективных методов выявления социальных связей между исследователями является кластер-анализ, основанный на сегментации тематических интересов.

Рекомендательная система подбора рецензентов фокусируется на поиске соответствующих рецензентов для научных документов. Рекомендательная система научных статей ориентирована на подбор

специализированных научных документов для исследователей. Цитатные рекомендательные системы, анализируя содержание основного текста, подбирают релевантные запросу цитаты [10].

Подходы, применяемые в современных рекомендательных системах для научных исследований, могут быть объединены в три группы [2]:

– подходы, основанные на анализе содержимого;

– методы коллаборативной фильтрации;

– гибридные методы.

Контент-ориентированные подходы

Контент-ориентированные подходы сосредотачивают внимание на сопоставлении текстовых документов с точки зрения близости ключевых слов и используют несколько методов, в том числе латентный семантический анализ (LCA) [5,9,13], векторная модель семантики (VSM) [0].

Контент-ориентированные методы для выработки рекомендаций используют информацию от самих объектов.

Для данных методов можно привести некоторые характеристики:

– предопределенные представление и организация документов;

– представление текущих интересов пользователя;

– наличие стадии сравнения, результатом которой является набор соответствующих документов;

– наличие стадии оценки выбранных документов;

– динамический характер интересов пользователя.

Интересы пользователя представляются в виде запросов, состоящих в большинстве случаев из ключевых слов, описывающих потребности пользователя.

В дополнение к указанным характеристикам следует отметить некоторые другие важные аспекты:

– выдача соответствующих документов может быть произведена как из статического корпуса, так и постоянно меняющегося корпуса;

– ранжирование документов может быть выполнено как по релевантности, так и по времени создания;

– запрос может быть сохранен в информационной модели пользователя.

Общими чертами всех моделей, используемых в контент-ориентированных рекомендательных системах, являются индексация и классификация содержимого каждого документа в корпусе документов. Приведем краткие описания трех моделей.

Модель логического поиска

В модели логического поиска пользовательский запрос может состоять из нескольких подзапросов (терминов), соединенных логическими операторами. Это модель «с точным соответствием», в которой терминам запроса должны соответствовать термины, найденные в соответствующих документах. Эта модель не предусматривает ранжирования релевантности.

Векторная модель

В векторной модели документ моделируются как вектор в многомерном векторном пространстве терминов [1]. Каждому измерению пространства соответствует термин из корпуса документов. Значение каждой из компонент вектора документа равно оценке *TFIDF* важности термина в тексте документа, которую можно определить следующим образом.

Обозначим *TF* (term frequency) – это нормализованная частота слова в тексте, которая определяется по формуле

$$TF(t,\ d) = \frac{\mathrm{freq}(t,\ d)}{\max\limits_{w \in D} \mathrm{freq}(w, d)}, \qquad (1)$$

где $\mathrm{freq}(t,\ d)$ – количество слов t в документе d. Величина *TF* принимает значения из отрезка [0,1].

Пусть $IDF(t,\ D)$ – обратная частота документов (inverse document frequency).

$$IDF(t,\ D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}, \qquad (2)$$

где $|D|$ – количество документов в наборе, $|\{d \in D : t \in d\}|$ – количество документов, в которых встречается слово t.

Искомая оценка *TFIDF* вычисляется как произведение *TF* на *IDF*.

$$TFIDF(t,\ d, D) = TF(t,\ d)\, IDF(t,\ D). \qquad (3)$$

Сходство между документами можно оценить с помощью вычисления косинуса угла между векторами.

Вероятностная модель

В вероятностных моделях в качестве меры релевантности запроса к различным документам используется вероятность. Для этого строится байесовский классификатор, который должен предсказать вероятность того, что страница p_j принадлежит к классу C_i (т.е. является важной или неважной) исходя из ключевых слов $k_{1;j}; ...; k_{n;j}$ на этой странице.

Методы коллаборативной фильтрации

Традиционные методы коллаборативной фильтрации, используемые в научных рекомендательных системах, точно предсказывают предметы интереса для активного пользователя, основываясь на ранее известных предпочтениях похожих пользователей. Более точно основное предположение формулируется следующим образом: пользователи, которые ранее имели похожие мнения по вопросам в некоторой предметной области, в будущем будут также иметь схожие мнения. В рекомендательных системах, основанных на коллаборативной фильтрации, организуется сбор мнений пользователей об объектах. Эта информация хранится в матрице рейтингов. Например, можно построить три различных рейтинговых матрицы: автор-цитирование, статья-цитирование, и цитирование-цитирование. Отметим, что часто возникает ситуация, когда матрица рейтингов оказывается разреженной. В этом случае значения нулевых элементов матрицы рейтингов активного пользователя заменяются совокупными рейтингами объектов, построенными на основании информации, полученной от других пользователей. Для этого в системе организуется поиск k пользователей, наиболее похожих на активного пользователя, и которых будем именовать соседями. Совокупные рейтинги соседей предполагается использовать в качестве рекомендаций для активного пользователя.

Существует несколько методов вычисления сходства между двумя пользователями [7,11]. Наиболее часто используется метод с использованием корреляции Пирсона. Для активного пользователя a и другого пользователя u корреляция Пирсона $w(a,u)$ определяется по формуле

$$w(a,u)=\frac{\sum_{k=0}^{n}\left(r_{ak}-\overline{r}_a\right)\left(r_{uk}-\overline{r}_u\right)}{\sigma_a\sigma_u}, \quad (4)$$

где суммирование выполняется по всем объектам с рейтингами от пользователей a и u; $\overline{r}_f$ – средний рейтинг пользователя f; σ_f – стандартное отклонение рейтинга пользователя f.

На основе сходства между всеми пользователями отбираются k наиболее похожих на активного пользователя с дальнейшим объединением их рейтингов. Для этого формируется набор элементов, у которых присутствуют рейтинги соседей и у которых отсутствуют рейтинги активного пользователя. Совокупный рейтинг p_{ai} для активного пользователя a и объекта i из полученного набора определяется по формуле

$$p_{ai}=\overline{r}_a+\frac{\sum_{u=1}^{k}\left(r_{ui}-\overline{r}_u\right)w(a,u)}{\sum_{u=1}^{k}w(a,u)}, \quad (5)$$

где суммирование выполняется по всем k соседям пользователя a.

Очевидно, что данные методы являются менее эффективными в случае недостаточного числа оценок от других пользователей.

Использование методов коллаборативной фильтрации эффективно для выработки рекомендаций по статьям, цитатам и при поиске экспертов.

Гибридный подход

Гибридный подход предусматривает сочетание контент-ориентированных методов и коллаборативной фильтрации. Комбинирование методов позволяет избежать ограничений, свойственных каждому подходу. Например, Хванг и Чжуан [8] предложили подход, сочетающий информацию о содержании статьи и информацию об интернет-активности её использования, для выработки рекомендаций в контексте цифровой библиотеки. Хе и др. [0] создали интегральную модель, комбинирующую лингвистическую модель с анализом цитируемости, для получения рекомендаций относительно цитат для научно-исследовательских работ.

Заключение

В данной работе была изложена классификация методов, используемых в системах научных рекомендаций. Каждый подход обладает своими преимуществами и ограничениями, учитывая которые определяются области их эффективного применения. Отмечено, что комбинирование различных алгоритмов позволяет построить более точную систему научных рекомендаций.

Работа выполнена при финансовой поддержке РФФИ (проект №16–29–12875).

Список литературы

1. Гомзин А.Г., Коршунов А.В. Системы рекомендаций: обзор современных подходов // Труды Института системного программирования РАН. – 2012. – Т. 22. – С. 401–417.

2. Городецкий В.И., Тушканова О.Н. Онтологии и персонификация профиля пользователя в рекомендующих системах третьего поколения // Онтология проектирования. – 2014. – № 3 (13). – С. 7–31.

3. Елизаров А.М., Жижченко А.Б., Жильцов Н.Г., Кириллович А.В., Липачёв Е.К. Онтологии математического знания и рекомендательная система для коллекций физико-математических документов // Докл. РАН, 467:4 (2016). – С. 392–395.

4. Biswas, H.K., Hasan, M. Using publications and domain knowledge to build research profiles: An application in automatic reviewer assignment. In Proceedings of the International Conference on Information and Communication Technology, ICICT'07. 2007. pp. 82–86. Bangladesh, Dhaka.

5. Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet allocation // Journal of Machine Learning Research. – 2003. – Vol.3. – pp. 993–1022.

6. He Q., Kifer D., Pei J., Mitra P., & Giles C.L. Citation recommendation without author supervision // Paper presented at the Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. – Hong Kong, 2011. – pp. 755–764.

7. Herlocker J.L. Understanding and Improving Automated Collaborative Filtering. Ph.D., University of Minnesota, 2000.

8. Hwang S., Chuang S. Combining article content and Web usage for literature recommendation in digital libraries // Online Information Review. – 2004. – 28(4). – pp. 260–272.

9. Manning C.D., Raghavan P., Schütze H. Introduction to information retrieval. – Cambridge: Cambridge university press, 2008. – Vol. 1. No. 1. – P. 496.

10. McNee S.M., Kapoor N., Konstan J.A. Don't Look Stupid: Avoiding Pitfalls when Recommending Research Papers. In Proceedings of the 2006 ACM Conference on Computer Supported Cooperative Work (CSCW 2006), Banff, Canada, November 2006, pp. 171–180.

11. Resnick P., Iacovou N., Suchak M., Bergstrom P., Riedl J. GroupLens: An Open Architecture for Collaborative Filtering of Netnews, in Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work, 1994. – pp 175–186.

12. Silva T.A profile-boosted research analytics framework to recommend journals for manuscripts // Journal of the Association for Information Science & Technology. – 2015. – 66(1). – pp. 180–200.

13. Steyvers M., Smyth P., Rosen-Zvi M., Griffiths T. Probabilistic author-topic models for information discovery // Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, Washington, August 22–25, 2004.