RAIF 2019

сессия Natural Language Processing

Knowledge Factory

МАСТЕРСКАЯ ЗНАНИЙ: ПОИСКОВО-РЕКОМЕНДАТЕЛЬНАЯ СИСТЕМА ДЛЯ СИСТЕМАТИЗАЦИИ ПРОФЕССИОНАЛЬНОГО КОНТЕНТА

КОНСТАНТИН ВОРОНЦОВ K.V.VORONTSOV@PHYSTECH.EDU

Лаборатория Машинного Интеллекта МФТИ http://mipt.ai/
Компания Digital Decisions (AITHEA) http://aithea.com/

КОНЦЕПЦИЯ «МАСТЕРСКОЙ ЗНАНИЙ»

«Огромное и все возрастающее богатство знаний разбросано сегодня по всему миру. Этих знаний, вероятно, было бы достаточно для решения всего громадного количества трудностей наших дней, но они рассеяны и неорганизованы. Нам необходима очистка мышления в ***своеобразной мастерской*, где можно получать, сортировать, суммировать, усваивать, разъяснять и сравнивать знания и идеи.**» – *Герберт Уэллс, 1940*

(An immense and ever-increasing wealth of knowledge is scattered about the world today; knowledge that would probably suffice to solve all the mighty difficulties of our age, but it is dispersed and unorganized. We need a sort of mental clearing house for the mind: **a depot where knowledge and ideas are received, sorted, summarized, digested, clarified and compared** – *Herbert Wells, 1940*)

Сегодня технологии IR-ML-NLP позволяют решить эту задачу.

ОТ ПОИСКА ИНФОРМАЦИИ К МАСТЕРСКОЙ ЗНАНИЙ

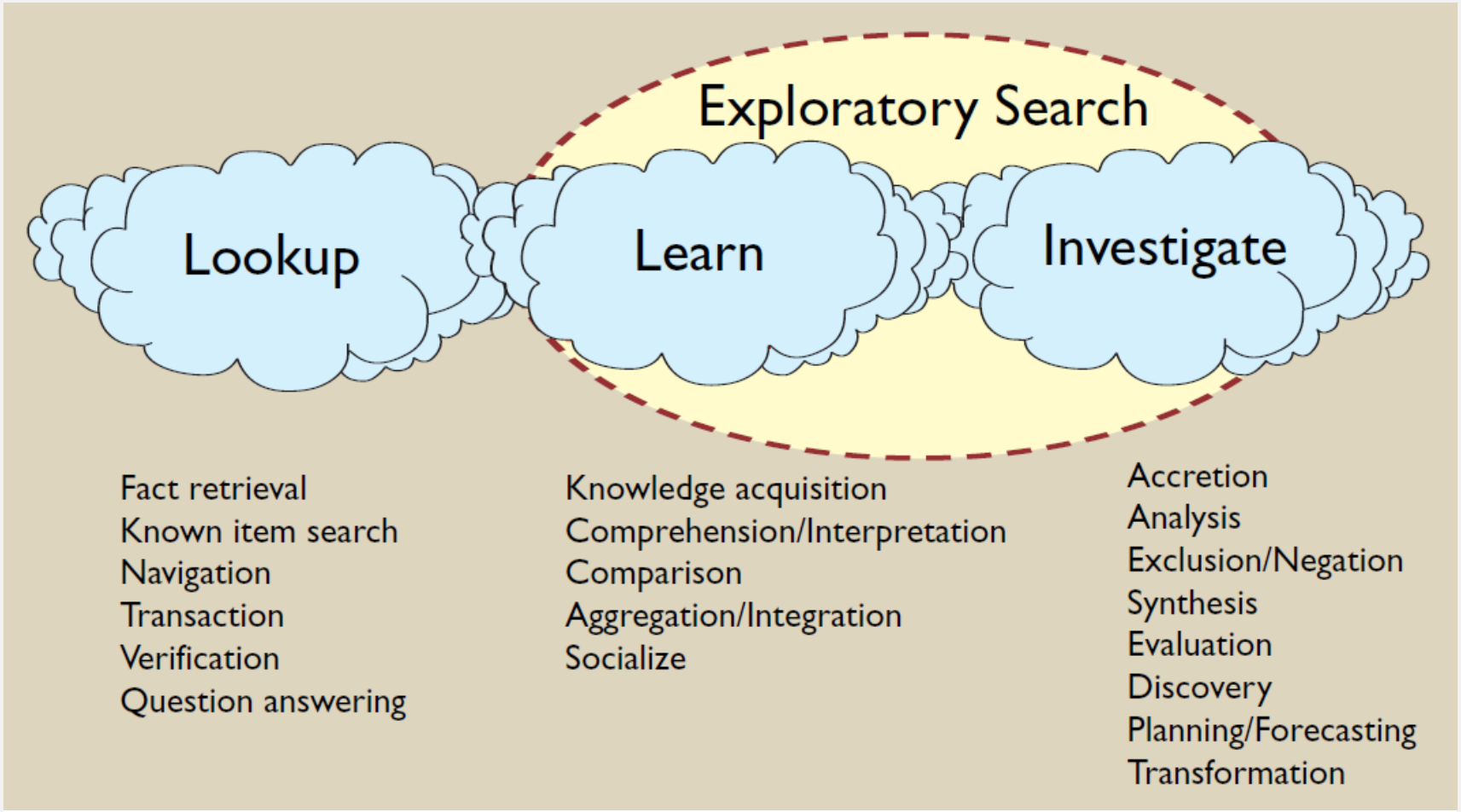
Обычный поиск:
«нашёл и забыл»

Google bing Yandex Baidu 百度

Мастерская знаний:

- ищу – чтобы накапливать
- накапливаю – чтобы анализировать
- анализирую – чтобы понимать
- понимаю – чтобы передавать

КОНЦЕПЦИЯ РАЗВЕДОЧНОГО ПОИСКА



Gary Marchionini. Exploratory Search: from finding to understanding. 2006.

ОСОБЕННОСТИ РАЗВЕДОЧНОГО ПОИСКА

1. An evolving search process
2. Several one-off pinpoint searches
3. An evolving information need
4. Multiple targets / goals of search
5. Multiple possible answers
6. Not an expected exact answer
7. A serendipitous attitude
8. An open-ended search activity
9. An anomalous state of knowledge
10. Multifaceted search
11. Uncertainty is fluctuating

#	Characteristics	Definition	References (non exhaustive list)
1	An evolving search process	The user adopts an opportunistic behavior, and will change or specify the objective or goals of search or even the strategies used to achieve them through multiple queries reformulation or refinement. During the search, the user can accomplish forward or backward steps.	[11, 13, 15, 22–24]
2	Several one-off pinpoint searches	Throughout the search session, the user can do several one-off pinpoint searches, e.g. she's looking for a specified information to better understand a result or the reason why it was proposed. These pinpoint searches can be related to the exploratory search task or not. This is closely related to sensemaking activities.	[10, 22, 23]
3	An evolving information need	Throughout the search session, the user has an evolving information need. The elements or results discovered may change her information need and the way she first considered the framework of the search. This evolution of the information need may appear several times in one search session. It is closely related to characteristic n°1.	[11, 22–24]
4	Multiple targets/ goals of search	The user may not have one single precise goal, but rather one vague objective and several smaller goals which may change or evolve during the exploratory search task so as to achieve it.	[2, 11, 12, 15, 22, 23]
5	Multiple possible answers	As the user has one vague objective and several smaller goals to achieve it (see characteristic n°3.), the user might not have one precise answer but an aggregate of relevant information which will help her go further in her reflection and exploratory search process.	[2, 11, 12, 20, 23]
6	Not an expected exact answer		
7	A serendipitous attitude	It is the faculty to be surprised and to pay attention to it. The user carries out her search by adopting a serendipitous attitude; with such open mindedness, she can allow herself to be surprised by one unexpected element. She then exploits this discovery by changing the search strategy or search goal/objective, etc.	[10, 20, 22]
8	An open ended search activity which can occur over time	The user might never end her exploratory search. She can stop it for multiple reasons (she considers she has enough information to perform another task for example; she doesn't have time to carry on the search; etc.), and she will continue the search few hours/days/weeks/months/years later.	[10, 12, 15, 22–24]
9	An Anomalous State of Knowledge (ASK) and an ill-structured (vague, general or unsure) context of search or goals	At the beginning, the user has an ASK and a general context of search: she knows the motivation to start the search, but does not have a precise idea of what she is actually looking for (type of results, kind of information). She only has a lack of knowledge, a vague objective of search but no specific of definitive plan to attain it.	[11, 15, 22–24]
10	Multifaceted	During the exploratory search, the user selects one or multiple filters or *facets*, to explore the information space. She will try to find an approach to her problem, she may find an angle of attack or a framework which may include these facets of the explored subject.	[2, 11, 22–24]
11	Uncertainty is fluctuating	The user starts the search with an intense feeling of uncertainty. The level of uncertainty is intrinsically linked to the specification of the problem. The further the user goes in her search tasks (she will specify her objective and maybe define an approximate plan), the more she reduces her uncertainty. But if somewhere along the way she changes her objectives, the uncertainty will tend to increase again.	[11, 23, 24]

E.Palagi et al. A Survey of Definitions and Models of Exploratory Search. 2017.

ОСОБЕННОСТИ РАЗВЕДОЧНОГО ПОИСКА:
РАЗВЕДОЧНЫЙ ПОИСК – ЭТО ПРОЦЕСС

1. An evolving search process

- разведочный поиск – это многошаговый процесс
- каждый шаг – это переформулировка или дополнение запроса

9. An anomalous state of knowledge

- в начале поиска у пользователя есть лишь мотивации,
- но нет знаний и нет определённого плана, как эти знания получать

4. Multiple targets / goals of search

- нет конкретной, точно определённой цели поиска
- есть лишь общий интерес и эволюционирующие подцели

E.Palagi et al. A Survey of Definitions and Models of Exploratory Search. 2017.

ОСОБЕННОСТИ РАЗВЕДОЧНОГО ПОИСКА:
НЕОПРЕДЕЛЁННОСТЬ В ПРОЦЕССЕ ПОИСКА

5. Multiple possible answers

- возможных правильных ответов может быть много

6. Not an expected exact answer

- не существует единственного ожидаемого правильного ответа

7. A serendipitous attitude

- любой шаг поиска может давать неожиданные новые знания

3. An evolving information need

- на любом шаге цели и стратегии поиска могут измениться

11. Uncertainty is fluctuating

- в процессе поиска растёт знание и уменьшается неопределённость
- но на любом шаге изменение цели может снова её увеличить

E.Palagi et al. A Survey of Definitions and Models of Exploratory Search. 2017.

ОСОБЕННОСТИ РАЗВЕДОЧНОГО ПОИСКА:
РАЗВЕТВЛЁННОСТЬ ПРОЦЕССА ПОИСКА

10. Multifaceted search

- в процессе поиска используются различные фильтры (фасеты)
- примеры: по авторам, по тематике, по свежести, по сложности

2. Several one-off pinpoint searches

- многократные точечные одноразовые ответвления поиска
- примеры: найти определение понятия, посмотреть первоисточник

8. An open-ended search activity which can occur over time

- процесс поиска никогда на заканчивается
- пользователь может возобновить поиск после длительного перерыва

E.Palagi et al. A Survey of Definitions and Models of Exploratory Search. 2017.

КТО И ЧТО ДЕЛАЕТ В МАСТЕРСКОЙ ЗНАНИЙ

Проектные группы – основной тип пользователей сервиса

Подборка документов – основной инструмент представления долгосрочного тематического интереса проектной группы.

Основные функции сервиса:

1. *Поисково-рекомендательные:*
   накопление и мониторинг знаний

2. *Аналитические:*
   суммаризация, систематизация и понимание знаний

3. *Коммуникативные*:
   представление и передача знаний

ПОИСКОВО-РЕКОМЕНДАТЕЛЬНЫЕ ФУНКЦИИ

1. Накопление знаний

- Используются только надёжные источники информации
- Всё, что нужно по моей теме, находится в моей подборке
- Могу искать документы по коротким текстовым запросам
- Могу получать рекомендации по подборке или по документу

2. Мониторинг знаний

- Подборка является постоянно действующим поисковым запросом
- Когда появляется новая рекомендация, система меня уведомляет

АНАЛИТИЧЕСКИЕ ФУНКЦИИ

3. Понимание знаний

- Система подсказывает мне, что читать в первую очередь

4. Суммаризация знаний

- Могу видеть основные идеи без прочтения документов целиком
- При написании обзоров по подборкам пользуюсь рекомендациями цитат, ссылок, фраз, вариантов продолжения

5. Систематизация знаний

- Могу разделять подборку на кластеры, «раскладывать по полочкам»
- Могу ранжировать подборку по тематике, свежести, сложности, обзорности, «хайповости», цитируемости, актуальности, качеству
- Могу разделять и объединять подборки

КОММУНИКАТИВНЫЕ ФУНКЦИИ

6. Передача знаний

- Подборку видит вся проектная группа
- Мы можем добавлять к статям рефераты, комментарии и теги
- Можем добавлять в подборку свои документы
- Можем открыть доступ к подборке

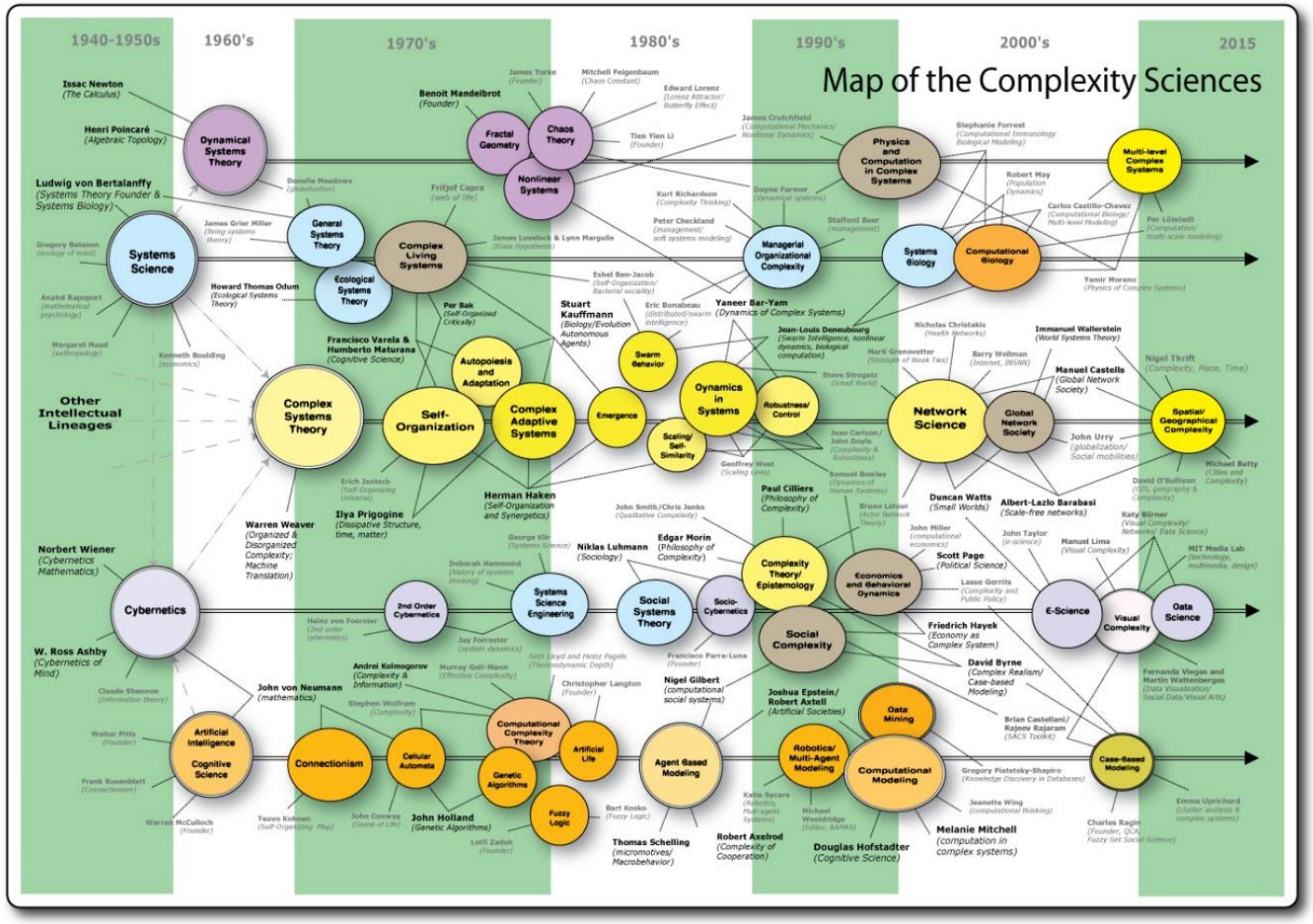
7. Представление знаний

- Интерактивные «карты знаний» и инфографика помогают видеть и объяснять ключевые идеи, содержащиеся в подборке
- Визуальное представление подборки, сделанное средствами сервиса, является интеллектуальным продуктом пользователя

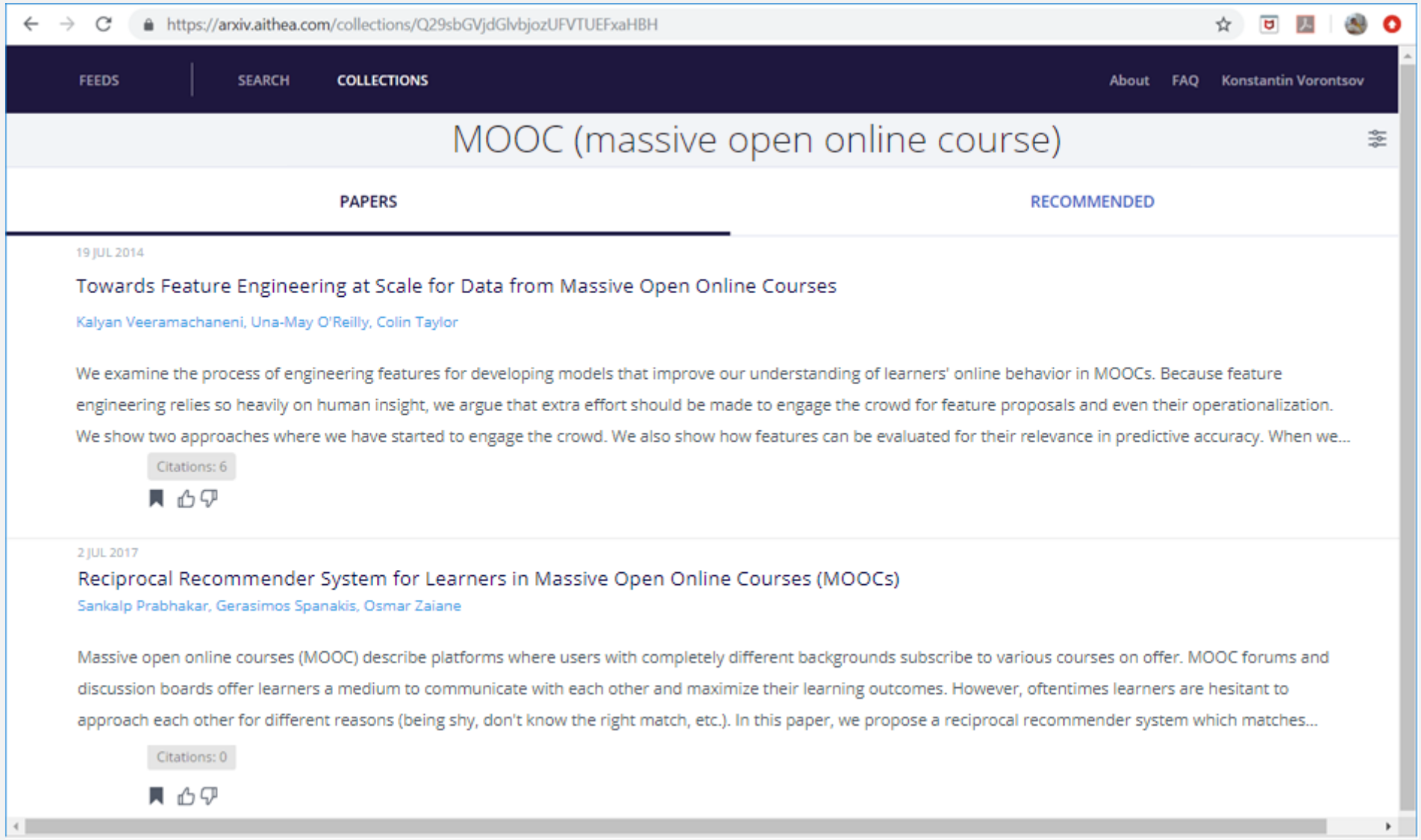
ПРИМЕР ВИЗУАЛИЗАЦИИ

Тематические карты в осях:

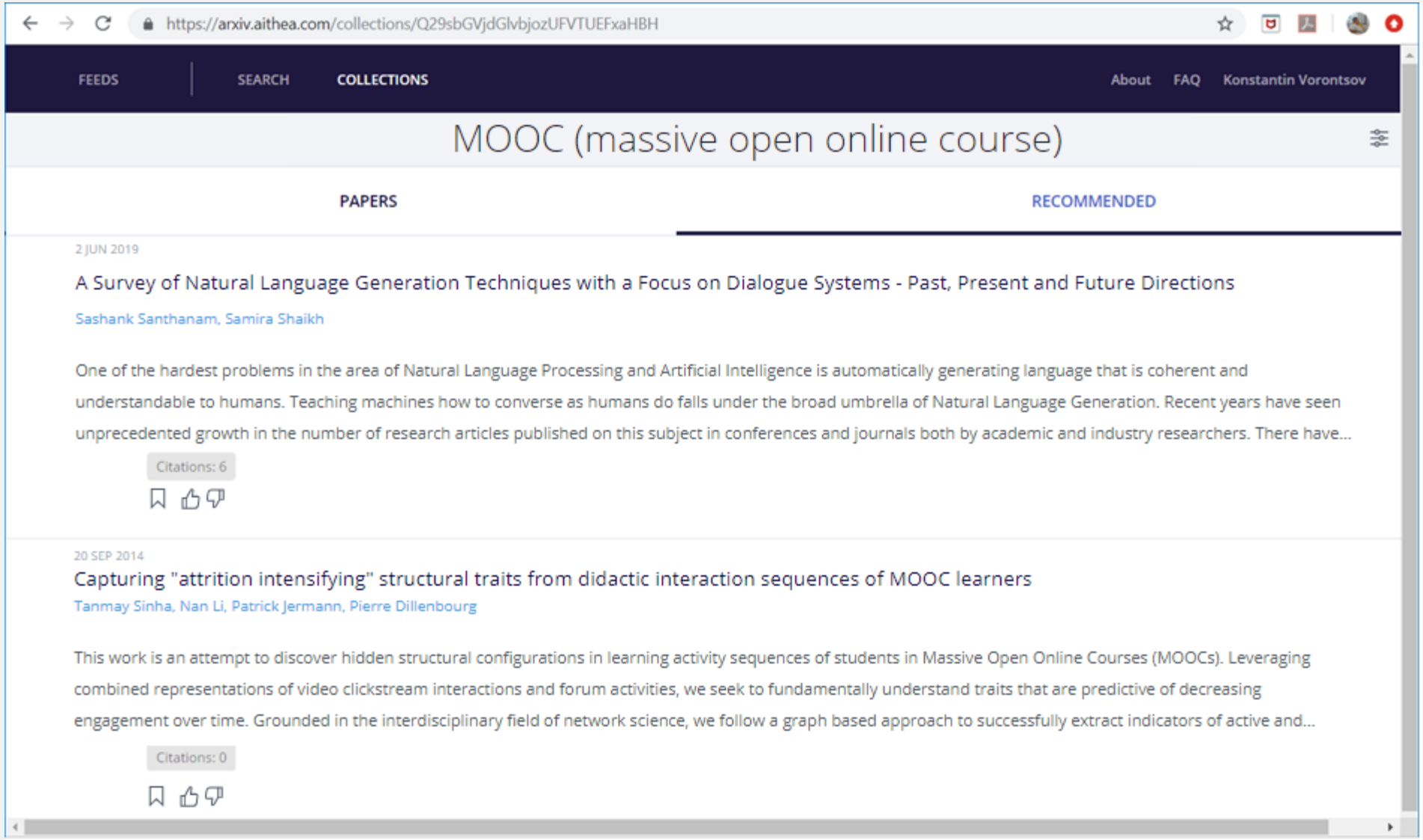
- время
- спектр тем
- сложность
- обзорность
- читабельность
- хайповость
- цитируемость



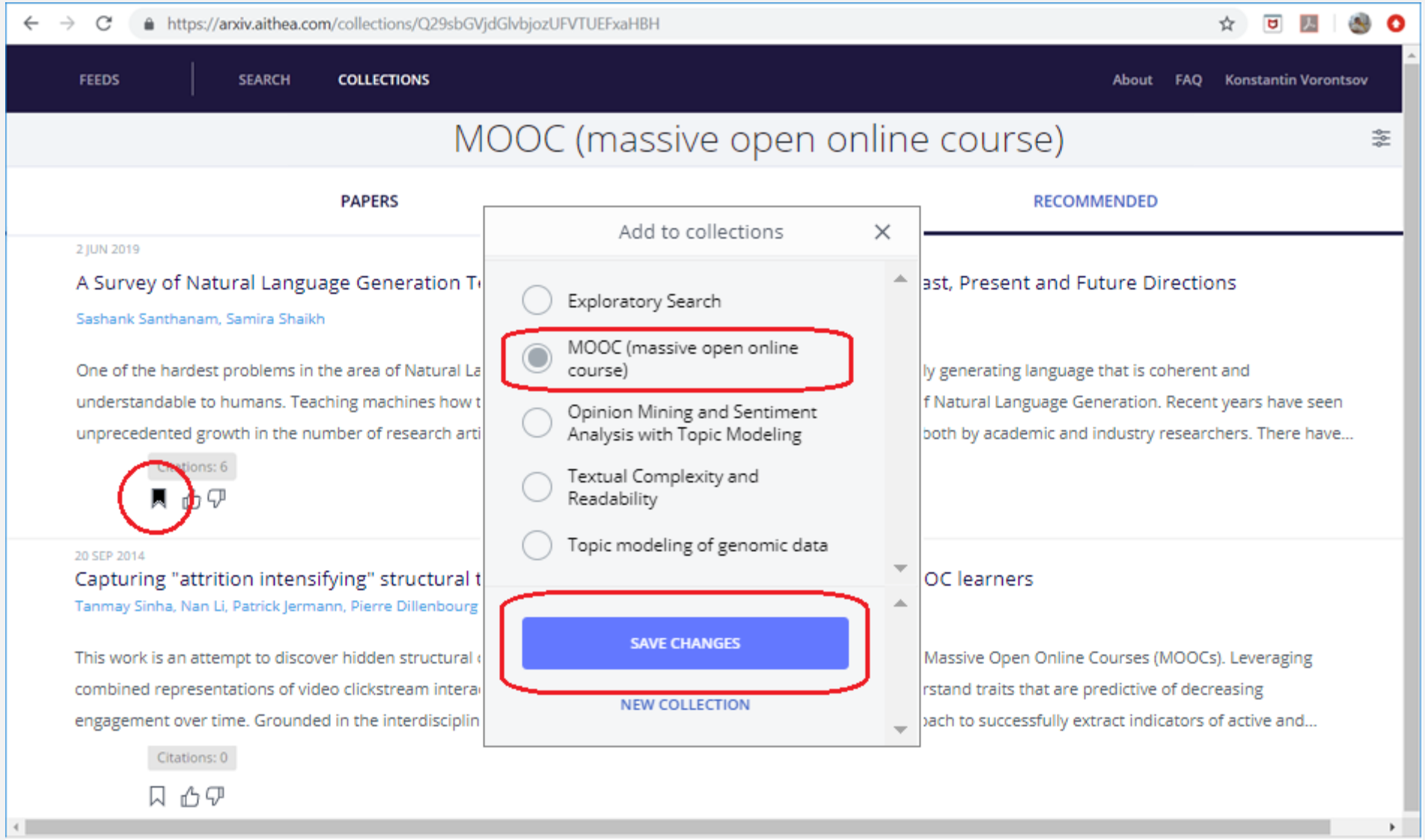
ПРОТОТИП: ARXIV.AITHEA.COM



ПРОТОТИП: ARXIV.AITHEA.COM



ПРОТОТИП: ARXIV.AITHEA.COM



ОДНА ПЛАТФОРМА – СЕМЕЙСТВО СЕРВИСОВ

Шаг 1 – прототип поисково-рекомендательного сервиса

arXiv.AITHEA.com – разведочный поиск по научным статьям

Шаг 2 – клонирование сервиса на других данных

core – по всем научным публикациям
mooc – по массовым открытым онлайн-курсам
pubmed – по биомедицинским исследованиям
popmed – по популярным медицинским статьям
popscience – по научно-популярным статьям
geeks – по технологическим блогам
patent – по патентным базам
legal – по законодательным и судебным актам
news – по новостным потокам

ОТЛИЧИЯ СЕРВИСОВ – В АНАЛИТИКЕ

Научные публикации

- **Подборка** по тематике исследований
- **Классификация** по подтемам, научным школам, терминологии
- **Выявление** трендов, направлений, подходов, открытых проблем

Массовые открытые онлайн-курсы

- **Подборка** по тематике курсов
- **Классификация** по уровню, пререквизитам,
- **Выявление** персональной образовательной траектории

Научно-популярный и просветительский контент

- **Подборка** по тематическим интересам пользователя
- **Классификация** по подтемам, возрасту целевой аудитории
- **Выявление** «точек входа» в науку, порядка чтения

ОТЛИЧИЯ СЕРВИСОВ – В АНАЛИТИКЕ

Отзывы и обзоры о потребительских товарах

- **Подборка** по назначению товара
- **Классификация** по потребительским свойствам товара
- **Выявление** различий в атрибутах товаров, брендах, продавцах

Новостные потоки

- **Подборка** по теме, проблеме или событию
- **Классификация** по тональности, акцентированию, умалчиванию
- **Выявление** полярных мнений и их источников

Акты арбитражных судов

- **Подборка** документов, схожих по существу дела
- **Классификация** по исходу дела
- **Выявление** наилучшей аргументации для суда

СУХОЙ ОСТАТОК

Мастерская знаний – концепция информационного поиска для профессиональных сообществ

arXiv.AITHEA.com – прототип разведочного поиска

Константин Воронцов K.V.Vorontsov@phystech.edu

Лаборатория Машинного Интеллекта МФТИ http://mipt.ai/

Компания Digital Decisions (AITHEA) http://aithea.com/

Machine Intelligence Laboratory

AITHEA