

Expert finding and expertise
retrieval

Поиск экспертов и извлечение
компетенций

Николай Карпов
НИУ ВШЭ Нижний Новгород
nkarпов@hse.ru

Поиск экспертов и извлечение компетенций

Задачи не имеют четкой постановки, так как существуют различные определения, что такое компетенции.

- В одних работах это область интересов человека (+ уровень компетентности в каждой)
- В других это навыки человека (что конкретно умеет делать, выражается отглагольным существительным)

Что часто понимают под компетенциями

- Область знания (управление рисками, формальная логика)
- Инструментальное средство (среда SPSS, пакет Matlab)
- Модель, теория, понятие (модель Эрроу-Дебре, дефлятор ВВП)
- Умение, навык (обработка древесины, разработка под iOS)

ИСТОЧНИКИ

- Balog, K and others: Expertise Retrieval, (2012). State-of-the-Art overview
- TREC Enterprise Track [Balog et al., 2008]
- Expert finding on DBLP data [Deng et al., 2008]
- Fang, H., Zhai, C.: Probabilistic models for expert finding. Advances in Information Retrieval. (2007).
- Serdyukov, P., Hiemstra, D.: Modeling documents as mixtures of persons for expert finding. (2008).
- Fomichov, V.: Semantics-Oriented Natural Language Processing (2009).
- Wei, X., Croft, W.B.: LDA-based document models for ad-hoc retrieval. (2006).
- Momtazi, S., Naumann, F.: Topic modeling for expert finding using latent Dirichlet allocation, (2013).
- Baroni, M., Lenci, A.: Distributional memory: A general framework for corpus-based semantics, (2010).
- Thomas L. Griffiths, Mark Steyvers: Finding scientific topics, (2004).
- Thomas Minka, John Lafferty: Expectation-propagation for the generative aspect model. (2002).

Поиск экспертов и извлечение компетенций

Поиск экспертов

- Дано: компетенции
- Найти: эксперта удовлетворяющего требованиям

Извлечение компетенций

- Дано: эксперт и результат его деятельности
- Найти: какими компетенциями обладает эксперт

Извлечение компетенций.

Приложения

- Системы управления компетенциями
 - Управления знаниями на предприятии
 - Составление профиля сотрудника
- Выбор рецензента для проекта или статьи
- Рекомендательные системы для выбора
 - работы
 - претендента
 - веб сайтов, блогов, статей

Извлечение компетенций – сложная задача



Извлечение компетенций

- Kivimki I., Panchenko A., Dessy A., Verdegem D., Francq P., Bersini H. and Saerens M. "A Graph-Based Approach to Skill Extraction from Text". In Proceedings of the 8th Workshop TextGraphs-8 Graph-based Methods for Natural Language Processing. EMNLP 2013: Conference on Empirical Methods in Natural Language Processing. Seattle, USA, October 18-21, 2013
- <http://aclweb.org/anthology/W/W13/W13-5011.pdf>
- Слайды Alexander Panchenko
www.slideshare.net/alexanderpanchenko/presentations

Извлечение компетенций

Цель работы системы:

- Сопоставить профессиональные компетенции с людьми на основе текстов, которые те пишут (электронная почта, блоги, форумы, статьи и так далее).

Инструменты:

- Список компетенций извлеченный из LinkedIn.
- Компетенции связанные ссылками со страницами Википедии.

Метод:

- 1 Найти страницу Википедии релевантную входному документу
- 2 Использовать активизацию широкой сети на сети ссылок Википедии, чтобы найти компетенции, близкие или центральные для релевантных страниц.

Извлечение компетенций системой Elisit

- Исследуется извлечение компетенций из текста, то есть ассоциация компетенций с текстовым документом.
- Что тут называется компетенциями? То, что называется «Skills» в системе LinkedIn
- Метод: Нахождение страницы Wikipedia релевантной профилю и Spreading activation на сети ссылок между страницами

Оценка работы системы

The image shows two side-by-side screenshots of the LinkedIn Skills & Expertise page for 'Machine Learning'. The left screenshot shows a list of related skills on the left side of the page, with a bracket and the number '20' indicating the count of professionals who have these skills. The right screenshot shows the same page but with a red box highlighting the 'Related Skills' section, which lists skills such as Feature Selection, Semi-supervised Learning, Classifier, Dimensionality Reduction, Graphical Models, Reinforcement Learning, Unsupervised Learning, Pattern Recognition, Text Classification, Recommender Systems, Natural Language Processing, Text Mining, Object Detection, Collaborative Filtering, SVM, Statistical Machine Translation, Mahout, Bayesian networks, and NLTK.

Skills & Expertise **Machine Learning** ▲11%

Primary Industry: Research
Machine learning, a branch of artificial intelligence, is a scientific discipline concerned with the design and development of algorithms that take as input empirical data, such as that from sensors or databases, and yield patterns or predictions thought to be features of the underlying mechanism that generated the data. A learner can take advantage of examples (data) to capture characteristics...
More on 'Machine Learning' at Wikipedia +

Machine Learning Professionals

June Leskovec
Assistant Professor at Stanford University
Ph.D. in Machine Learning, Machine Learning Department, School of Computer Science, Carnegie Mellon University, ...

Evgeniy Gabrilovich
Senior Staff Research Scientist at Google
Senior Staff Research Scientist at Google. 50+ publications in top international venues, 20+ patents filed...

Monica Rogati
VP of Data at Jambone
Excited to join Jambone to inspire people to be healthier - with a data twist. Powered data driven products...

Daphne Koller
co-founder at Coursera, Inc.

Ron Brachman
Chief Scientist and Head, Yahoo! Labs at Yahoo!

Paul Viola
Chief Scientist at Highpat, Inc.
Manager, Architect, Developer, and Researcher that is equally comfortable with advanced machine learning, web...

Related Companies

Google
Internet, San Francisco Bay Area
Follow

Carnegie Mellon University
Higher Education, Greater Pittsburgh Area
Follow

Microsoft
Computer Software, Greater Seattle Area
Follow

Amazon
Internet, Greater Seattle Area
Follow

Stanford University
Higher Education, San Francisco Bay Area
Follow

More companies...

Related Locations

Stanford, Santa Clara

Mountain View, Santa Clara

Palo Alto, Santa Clara

Cambridge, Middlesex

Redmond, King

* Based on LinkedIn data

More locations...

Search Skills & Expertise

Machine Learning ▲11%

Primary Industry: Research
Machine learning, a branch of artificial intelligence, is a scientific discipline concerned with the design and development of algorithms that take as input empirical data, such as that from sensors or databases, and yield patterns or predictions thought to be features of the underlying mechanism that generated the data. A learner can take advantage of examples (data) to capture characteristics...
More on 'Machine Learning' at Wikipedia +

Machine Learning Professionals

June Leskovec
Assistant Professor at Stanford University
Ph.D. in Machine Learning, Machine Learning Department, School of Computer Science, Carnegie Mellon University, ...

Evgeniy Gabrilovich
Senior Staff Research Scientist at Google
Senior Staff Research Scientist at Google. 50+ publications in top international venues, 20+ patents filed...

Monica Rogati
VP of Data at Jambone
Excited to join Jambone to inspire people to be healthier - with a data twist. Powered data driven products...

Daphne Koller
co-founder at Coursera, Inc.

Ron Brachman
Chief Scientist and Head, Yahoo! Labs at Yahoo!

Paul Viola
Chief Scientist at Highpat, Inc.
Manager, Architect, Developer, and Researcher that is equally comfortable with advanced machine learning, web...

Related Skills

Feature Selection

Semi-supervised Learning

Classifier

Dimensionality Reduction

Graphical Models

Reinforcement Learning

Unsupervised Learning

Pattern Recognition

Text Classification

Recommender Systems

Natural Language Processing

Text Mining

Object Detection

Collaborative Filtering

SVM

Statistical Machine Translation

Mahout

Bayesian networks

NLTK

Natural Language

Производится оценка того, на сколько хорошо система находит компетенции, отмеченные в LinkedIn

Оценка работы системы

VSM	Pre@5	Pre@10	R-Pre	Rec@100
TF-IDF	0.231	0.214	0.190	0.516
LogEntropy	0.216	0.212	0.193	0.525
LogEnt + LSA	0.180	0.181	0.163	0.491
LogEnt + LDA	0.193	0.174	0.159	0.470

Например, если брать топ 5 наиболее часто активируемых компетенций (из 27000) встречаются 1-2 релевантные компетенции из ≤ 20 отмеченных.

Поиск экспертов



Человек может сам может не знать до конца своих способностей

Профилеориентированный метод

- Формируется профиль эксперта, объединяющий все написанные им тексты
- По профилю строится языковая модель персоны
- Кандидат представляется в виде многомерной функции распределения терминов в словаре.
- По входному запросу определяется наиболее вероятная модель персоны, для генерации запроса

Поиск экспертов

Candidate Generation Models

$$P(e/Q) = \sum_d P(e, d/Q) = \sum_d P(e/d, Q)P(d/Q)$$

$P(d|q)$ – вероятность на сколько документ d релевантен запросу q

$$P(d/Q) = \frac{P(Q/d)P(d)}{P(Q)} \propto P(Q/d) = \prod_{q \in Q} P(q/d)^{n(q,Q)}$$

$$P(e|D) = \frac{a(e, D)}{\sum_{i=1}^m a(e_i, D)}, \quad (3)$$

Использование семантического анализа для поиска специалистов

- Semantics-Oriented Natural Language Processin. Vladimir A. Fomichov (2012)
- Usage of Semantic Analysis of Texts for Finding Specialists with Required Competencies. Igor V. Zakhlebin (2014)
- Используется профиле-ориентированный подход

Использование семантического анализа для поиска специалистов

- Предложен метод семантического поиска специалистов по набору составленных ими текстов
- В систему загружаются тексты: анкеты, резюме, проф. переписка, статьи и т.п.
- Для поиска пользователь вводит запрос определенной структуры (прил + сущ, сущ + сущ,)
- Система ищет специалистов, у которых в связанных с ними текстах присутствуют релевантные словосочетания. Чем большему числу критериев удовлетворяет специалист, тем выше он располагается в ранжировании.

Построение семантического представления (СП)

- Выделение морфологических признаков и лексемы
- К существительным применяется лексико-семантический словарь
- По начальной форме сопоставляются семантические значения (sem) и набор характеристик или сортов (st)

Lec	Sem	St_1 ... St_k
-----	-----	---------------

- К существительным применяется семантико-синтаксические шаблоны. Prep – предлог, Grc – падеж Rel – отношение.

Prep	St_1	St_2	Grc	Rel
------	------	------	-----	-----

В результате выполнения алгоритма формируется СП фрагмента текста – ориентированное дерево, в вершинах которых находятся Sem и ребра заданы Rel. (Триплеты Sem Rel Sem)

Пример построения семантического представления (СП)

lec	sem	sr2
Программист	программирование	ints, progr
Разработчик	программирование	ints, progr
Разработка	программирование	progr

Prep	sr1	sr2	grc	rel	ex
-	tool	progr	1	Сфера	С++ программист
под	progr	tool	1	Сфера	Программирование на С++

Документоориентированный метод

- Входной запрос сравнивается сначала с документом, а через него ассоциируется с автором
- Формируем набор признаков для документа
- Новый объект классифицируем по методу ближайшего соседа (к соседей)
- При этом признаки документов могут быть всевозможными:
 - TF-IDF
 - LogEntropy
 - LSA
 - LDA

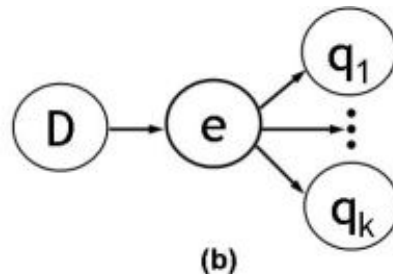
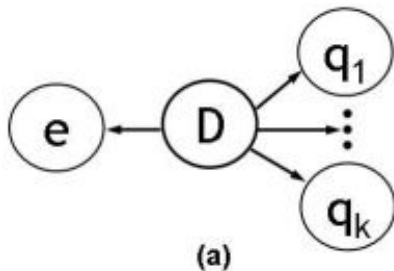
Документо-ориентированный метод

$$P(e, q_1, \dots, q_k) = \sum_{D \in R} P(D)P(e, q_1, \dots, q_k|D) \quad (1)$$

$$P(e, q_1, \dots, q_k|D) = P(e|D) \prod_{i=1}^k P(q_i|D) \quad (2)$$

$$P(e|D) = \frac{a(e, D)}{\sum_{i=1}^m a(e_i, D)}, \quad (3)$$

$$P(w|D) = (1 - \lambda_G) \frac{c(w, D)}{|D|} + \lambda_G P(w|G), \quad (4)$$



Person-Centric Expert Finding

- Человеко-ориентированный метод может быть рассмотрен как гибридный метод, объединяющий параметры документо-ориентированного и профиле-ориентированного метода.
- Ключевое допущение состоит в том, что уровень экспертизы может быть определен как совокупность ранжированных документов относящихся к персоне.

Поиск экспертов на основе скрытых ТОПИКОВ

Цель: поиск экспертов для

- формирования проектных команд
- рецензирования проектов и статей
- Экспертных оценок и комментариев

Методология: Topic modeling for expert finding using latent Dirichlet allocation.

Saeedeh Momtazi and Felix Naumann (2013)



Поиск экспертов на основе скрытых ТОПИКОВ

LDA модель

- Распределение вероятности слов по топикам:

$$P(w_i / z_k); i \in \overline{1, |\mathbf{W}|}, k \in \overline{1, |\mathbf{Z}|}$$

- Распределение вероятностей топиков по документам в коллекции

$$P(z_k / d_n); k \in \overline{1, |\mathbf{Z}|}, n \in \overline{1, |\mathbf{D}|}$$

- Идея метода состоит в том, чтобы рассматривать экспертов не отдельно от вероятностной модели LDA, а непосредственно внутри ее, так как имена экспертов это тоже слова

$$w_{i=ENames} = \mathbf{E} \Rightarrow \exists P(\mathbf{E} / \mathbf{Z})$$

Поиск экспертов на основе скрытых ТОПИКОВ

- Запрос Q обозначим как d_0 - новый документ, Используя обученную модель LDA можем построить для него распределение вероятностей по топикам

$$P(d_0 / E) = \sum_{z \in Z} P(d_0 / z, E) P(z / E)$$

$$P(d_0 / z, E) = P(d_0 / z, C) = P(d_0 / z) = \frac{P(z / d_0) P(d_0)}{P(z)} \propto P(z / d_0)$$

$$P(d_0 / E) \propto \sum_{z \in Z} P(z / d_0) P(E / z)$$

$$e_{\max} = \arg \max_{e \in E} (P(d_0 / e))$$

Применение модели

- Оценка точности работы алгоритма – порядка 0.3 на основе базы TREC 08
- Для русского языка апробация с использованием корпоративной базы публикаций сотрудников НИУ ВШЭ

The screenshot displays a search interface with the following components:

- Search Bar:** A blue button labeled "Искать" (Search).
- Key Words Section:** A text input field containing "введи те слова", a blue "Добавить" (Add) button, and a dropdown menu for "выбери тип поиска" (Select search type) with "распознавание образов" (Image recognition) selected. A blue "Удалить" (Remove) button is below.
- Taxonomy Section:** A dropdown menu for "Таксономия e library" (Library taxonomy) with "ОБЩИЕ ВОПРОСЫ" (General questions) selected. A blue "Добавить" (Add) button and a "выбери тип поиска" dropdown menu are below. A blue "Удалить" (Remove) button is at the bottom.
- Search Level Section:** A dropdown menu for "Уровень поиска" (Search level) with "Никакого/удалить фильтр" (No/Remove filter) selected. A blue "Искать" (Search) button is below.
- Results Section:** A blue button labeled "Искать" (Search) at the top. The main heading is "Результаты последнего поиска" (Results of the last search). A list of results shows "Савченко Андрей Владимирович" (Savchenko Andrei Vladimirovich) repeated. Below the list are blue buttons for "Детали" (Details) and "Отчет" (Report).
- Author Information Section:** A heading "Детальная информация" (Detailed information). It includes a profile picture of a man, a name "Савченко Андрей Владимирович", and affiliation "Кафедра информационных систем и технологий" (Department of Information Systems and Technologies).
- Classifiers Section:** A heading "Классификаторы" (Classifiers) with an empty text input field and a dropdown arrow.
- Keywords Section:** A heading "Ключевые слова научных интересов" (Keywords of scientific interests) with a text input field containing "критерии локального информационного распределения" and "принцип локального информационного распределения".

Спасибо за внимание!

nkarpov@hse.ru