8.3. Визуализация тематических моделей

Тематические модели, как правило, создаются для упрощения понимания и обеспечения навигации по большим текстовым коллекциям, поэтому важной задачей является визуализация тематических моделей. В последние годы было создано достаточно много средств визуализации, многие из которых находятся в свободном доступе. Большинство из этих инструментов ориентированы на то, чтобы визуализировать текстовые коллекции через web-интерфейсы.

8.3.1. Система TMVE

Система Topic Model Visualization Engine является одним из канонических примеров тематического навигатора с web-интерфейсом.

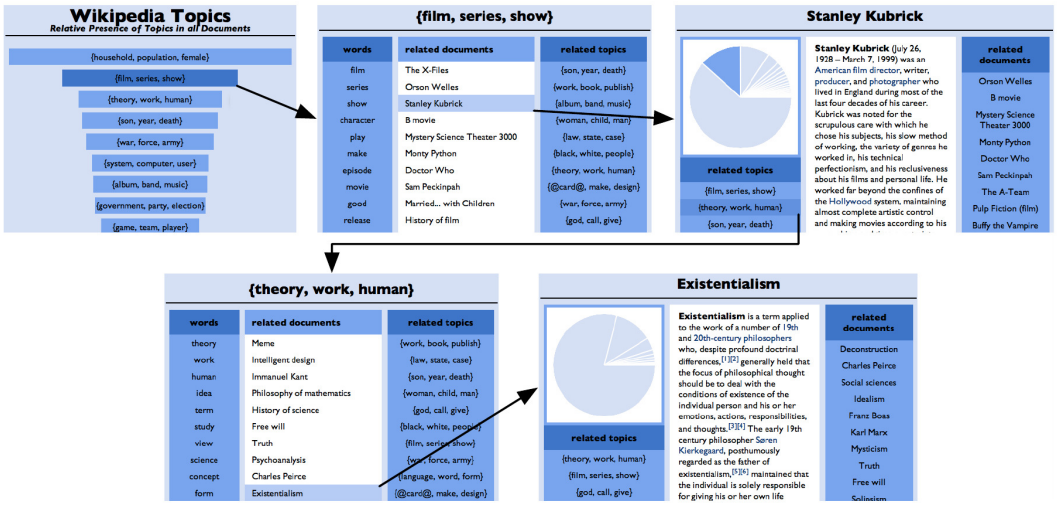


Рис. 8.1: «Wikipedia Topics», демонстрационный пример, представленный авторами TMVE.

На главной странице системы находится список тем, по каждой теме можно просмотреть документы и термины этой темы. Таким образом реализуется возможность навигации пользователя по коллекции.

8.3.2. Система TERMITE

Система TERMITE позволяет интерактивно визуализировать матрицу Φ и больше подходит для разработчиков тематической модели.

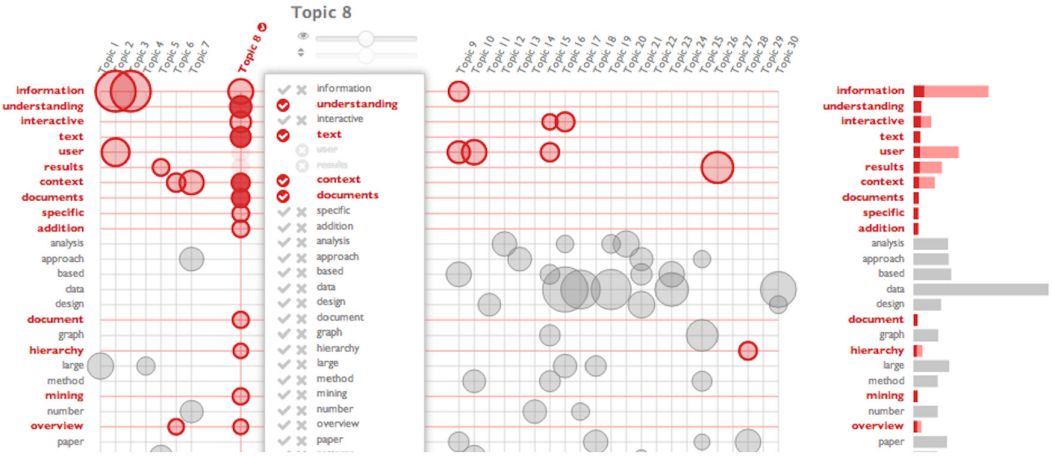


Рис. 8.2: Система TERMITE

8.3.3. Динамические модели, учитывающие время

Есть огромное количество средств визуализации для тематических моделей потоков новостей, научных статей или любых других коллекций, где каждому документу приписывается метка времени. Тогда строить тематические модели очень удобно, визуализируя их в виде графиков, на которых отображено, как развивались темы во времени, в какие моменты темы набирали популярность.

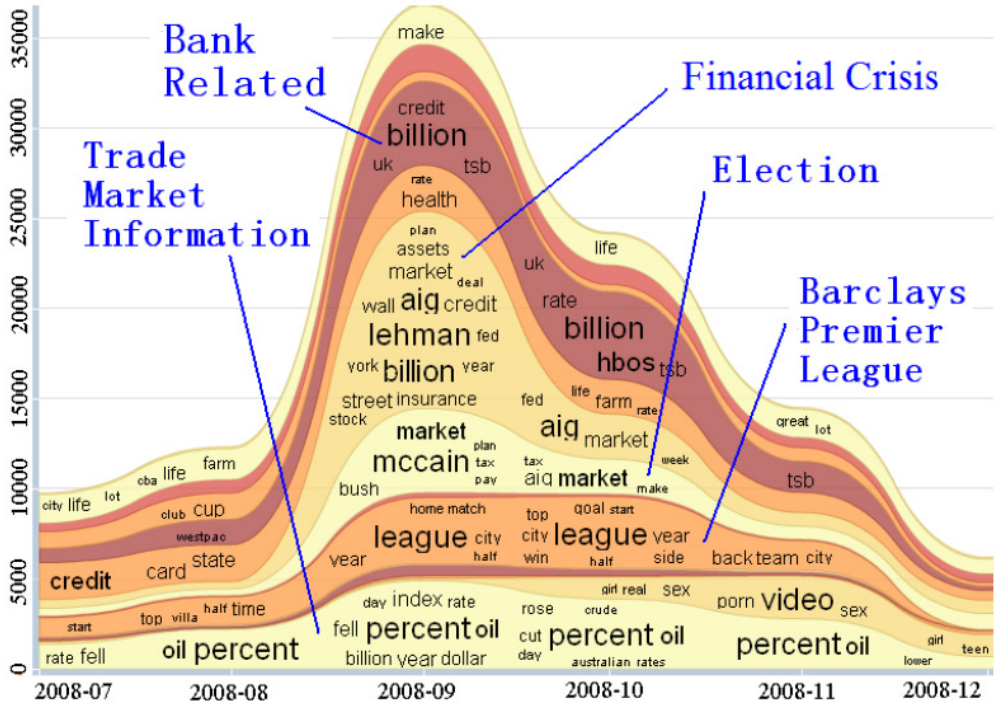


Рис. 8.3: На этом графике видны темы, которые возникли в связи с финансовым кризисом 2008.

На таких графиках можно изучать предвестники, последствия и связанные темы.

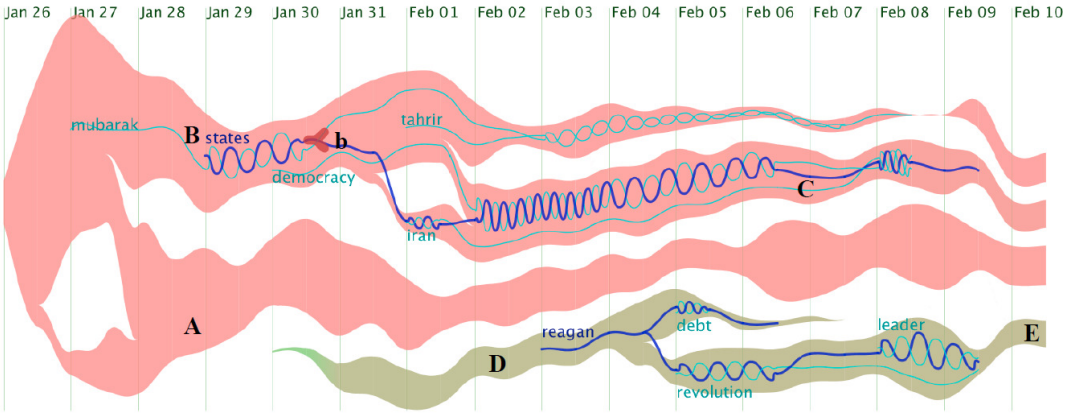
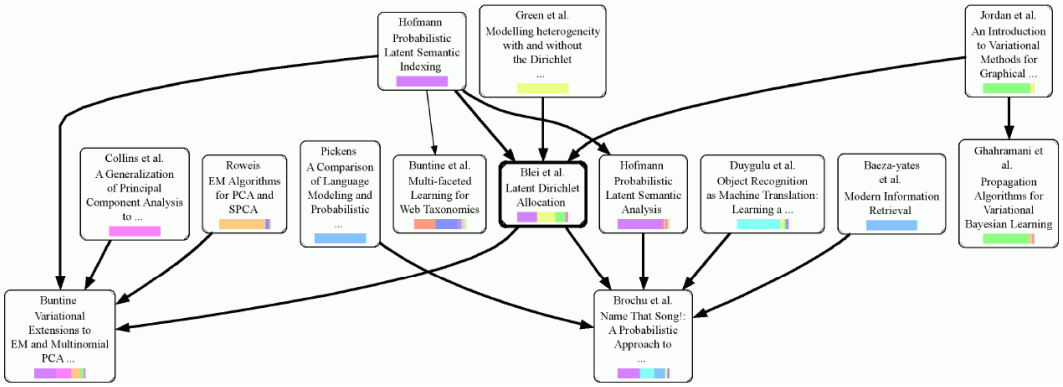


Рис. 8.4: Визуализация «река тем»: изображены моменты зарождения тем, исчезновения. Волнами изображаются траектории отдельных слов, причем часто колеблющаяся волна значит, что слово использовалось часто.

8.3.4. Динамические модели, учитывающие ссылки

Если тематическая модель учитывает не только слова, но также связи между документами, например связи цитирования между научными статьями, то можно ставить очень интересные задачи.

Например, можно попытаться ответить на вопрос, какие предшествующие работы действительно существенно повлияли на данную статью. В статье часто десятки ссылок, многие из которых чисто формальные, или дань вежливости, или же незначительные моменты, которые для данной статьи не важны. Оказывается, что это можно сделать с помощью тематической модели: выявить тематику статьи и выбрать из списка литературы те статьи, которые тоже соответствуют этой тематике.



С другой стороны, использование ссылок и цитат позволяет уточнить саму тематическую модель. Для этого предполагается, что если статья ссылается на другую, то у них есть общая тематика, и это учитывается с помощью регуляризатора.

8.3.5. Выявление взаимосвязей между темами

Оказывается, что можно выявлять связи между темами. Это особенно хорошо получается на коллекциях научных текстов. Так, например, в статье про археологию скорее появится термин из геологии, чем из генетики. Выявление таких связей между отраслями знаний представляет отдельный прикладной интерес.

Если каждую тему изображать в виде вершины графа, а ребро проводить только в том случае, когда соответствующие две темы часто появлялись в документах одновременно, то получившаяся тематическая модель будет называться кореллированной тематической моделью.

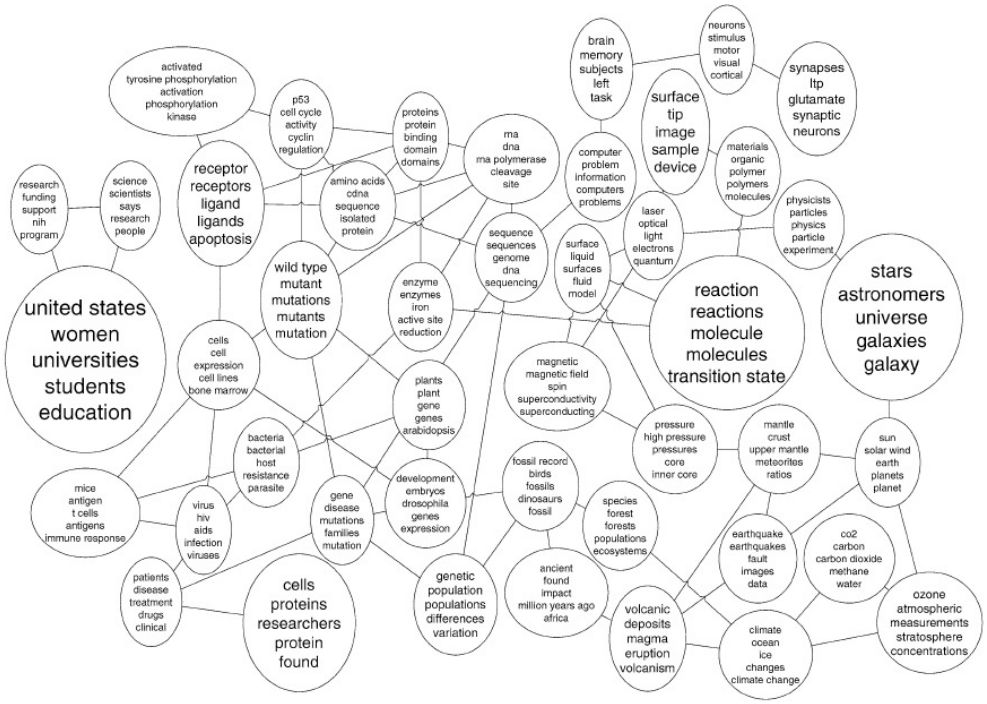


Рис. 8.5: Кореллированная тематическая модель, построенная на текстах из журнала Science.

Еще одно применение тематических моделей — использование в поисковых машинах. В частности, они используются для выбора документов, которые потенциально могут быть релевантны запросу.

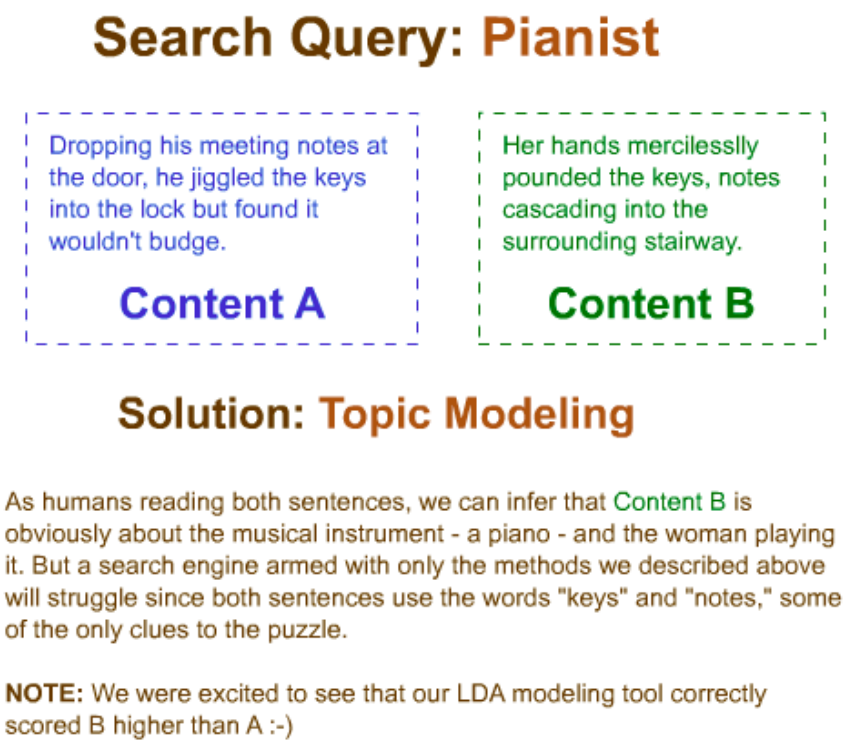


Рис. 8.7: Тематические модели позволяют отличить релевантный документ от нерелевантного за счет построения тематического профиля.

8.4.3. Методы тематического моделирования

Одним из самых первых подходов к построению тематических моделей, Probabilistic Latent Semantic Analysis (PLSA), был предложен в 1999 году. В этом методе ставилась задача матричного разложения с дополнительными условиями на нормировку столбцов матриц (столбцы матриц должны представлять собой вероятностное распределение), которая решалась методом максимального правдоподобия, стандартным методом в статистике.

В 2003 году эта же задача была рассмотрена в байесовской постановке, в которой вместо матриц строится вероятностное распределение над матрицами. Этот подход получил название Latent Dirichlet Allocation (LDA), или латентное размещение Дирихле. Стоит отметить, что LDA — самая изученная модель тематического моделирования.

Не так давно был разработан другой подход, Additive Regularization of Topic Models (ARTM), который заключался в регуляризации PLSA с целью получения лучших моделей. Для этого предполагается ввести дополнительные критерии как регуляризаторы в модель PLSA, за счет чего модель получается более гибкой и ее можно адаптировать к большему числу задач.

LDA	**ARTM**
Очень популярный	Молодой
Множество модификации для различных задач	Мощный аппарат регуляризаторов для модифицирования модели
Для каждого усложнения нужно искать реализацию	Одна реализация для разных задач
Нужно настраивать гиперпараметры	Нужно настраивать параметры регуляризации