

СОДЕРЖАНИЕ

Стр.

Введение..................................................................................................... 3

Актуальность работы……………………………………………. 3

Цель и задачи работы…………………………………………….3

1.Обзор основных подходов..................................................................... 4

1.1. Старые подходы.......................................... 4

1.2. Современные подходы .......................................... 7

2. Выявление лучших для использования на практике........................... 10

Заключение ................................................................................................. 11

Результаты работы……………………………………………….. 11

Список использованных источников........................................................ 11

Введение

Актуальность работы

В связи с ростом популярности социальных сетей растет потребность в системах, которые могут извлекать полезную информацию из

большого объема данных. Одна из задач – это выявления тем общения среди пользователей и дальнейшее наблюдение за развитием темы. Это позволит быстро узнать основные темы в реальном времени, которые волнуют людей, тратя при этом минимальное количество времени.

Цели и задачи работы

Целью данного обзора является изучение подходов к проблеме отслеживания эволюции тем. Для достижения этой цели, мы поставим следующие задачи:

1. Изучить и сравнить существующие методы и модели, используемые для отслеживания эволюции тем.
2. Выявить преимущества и недостатки каждого подхода.
3. Определить лучшие методы для программной реализации.
4. **Обзор основных подходов**
   1. Старые подходы
5. **Dynamic Topic Models**

**Авторы: David M. Blei и John D. Lafferty**

Блей и Лафферти разработали динамическую тематическую модель (DTM), предположив, что тематические модели развиваются постепенно во времени и распределяются нормально. В частности, анализ LDA по k компонентам проводится на каждом временном отрезке t. Каждая тема моделируется как гауссовский процесс, центрированный на предыдущем значении.

**Преимущества:**

* DTM может обнаруживать и отслеживать изменения в темах с течением времени, что может быть полезно для анализа больших коллекций документов, таких как новостные архивы или научные статьи.

**Недостатки:**

* DTM может быть вычислительно сложным, особенно для больших коллекций документов.
* Модель предполагает, что темы меняются плавно с течением времени, что может не всегда быть верно. Некоторые темы могут появляться и исчезать внезапно.

1. **Streaming First Story Detection with application to Twitter**

**Авторы: Sasa Petrovic, Miles Osborne, Victor Lavrenko**

Авторы представляют алгоритм, который может обнаруживать новые истории в потоке твитов в реальном времени. Этот метод использует кластеризацию для группировки похожих твитов вместе и определения новых историй.

**Преимущества:**

* Этот подход может быть очень полезен для обнаружения новых событий или тем в реальном времени на платформах социальных сетей, таких как Twitter.
* Метод может быть масштабируемым и эффективным, поскольку он способен обрабатывать большие объемы данных в реальном времени.

**Недостатки:**

* Метод может быть чувствителен к шуму и нерелевантным твитам, которые могут влиять на точность обнаружения историй.
* Этот подход может требовать значительных вычислительных ресурсов для обработки больших потоков данных.
* Метод может столкнуться с проблемами при обнаружении историй, которые медленно развиваются или меняются с течением времени.

1. **PET: A Statistical Model for Popular Events Tracking in Social Communities**

**Авторы: Cindy Xide Lin, Bo Zhao, Qiaozhu Mei, Jiawei Han**

В статье представлена статистическая модель для отслеживания популярных событий в социальных сообществах. Модель анализирует поведение пользователей и их взаимодействие с контентом для определения популярности событий. PET моделирует популярность событий со временем, учитывая всплески интереса пользователей, распространение информации в структуре сети и эволюцию текстовых тем. Используется модель Гиббса для моделирования влияния исторического статуса и зависимостей в графе.

Преимущества:

* Модель PET может быть полезна для анализа тенденций и популярности событий в социальных сообществах.
* Этот подход может помочь в понимании динамики социальных сообществ и поведения пользователей.

Недостатки:

Те же недостатки, что и предыдущего подходы.

**Сравнение PET с другими работами:**

* В отличие от некоторых других подходов, таких как “Streaming First Story Detection with application to Twitter”, модель PET фокусируется на отслеживании популярности событий, а не просто на обнаружении новых историй или событий.
* В отличие от “Dynamic Topic Models”, модель PET не только отслеживает изменение тем со временем, но и учитывает популярность событий и взаимодействие пользователей с контентом.
* В отличие от “Detecting Topic Evolution in Scientific Literature: How Can Citations Help?”, модель PET фокусируется на социальных сообществах и анализе поведения пользователей, а не на анализе цитирований в научной литературе.

1. **Generating Event Storylines from Microblogs**

**Автор: Chen Lin, Chun Lin, Jingxuan Li**

Статья “Generating Event Storylines from Microblogs” исследует проблему создания сюжетных линий из микроблогов для пользовательских запросов. Эта задача является сложной из-за разреженной, динамической и социальной природы микроблогов.

**Главные нововведения:**

* Авторы предлагают двухуровневое решение: сначала они используют языковую модель с динамической псевдо-обратной связью для получения релевантных твитов, а затем генерируют сюжетные линии через оптимизацию графа.
* Проблема генерации сюжетных линий на полученных микроблогах формулируется как задача оптимизации на основе графа и решается с помощью приближенных алгоритмов минимального весового доминирующего набора и направленного дерева Штейнера.

**Преимущества:**

* Эксперименты на наборах данных Twitter демонстрируют эффективность предложенных методов на каждом уровне и в целом по всей системе.
* Генерация сюжетных линий из микроблогов может быть полезна для промышленности, академического сообщества и конечных пользователей.

**Недостатки:**

* Из-за разреженной, динамической и социальной природы микроблогов, такой подход может столкнуться с проблемами, связанными с шумом в данных, изменчивостью тем и сложностью отслеживания социальных взаимодействий.

**Сравнение “Generating Event Storylines from Microblogs” с подходом PET:**

* Оба подхода фокусируются на взаимодействии между текстовым содержимым и социальными сетями.
* Оба подхода учитывают всплески интереса пользователей и эволюцию текстовых тем.
* В отличие от “Generating Event Storylines from Microblogs”, PET также учитывает распространение информации в структуре сети.
* В то время как “Generating Event Storylines from Microblogs” использует двухуровневое решение с оптимизацией графа для генерации сюжетных линий, PET использует модель Гиббса для моделирования влияния исторического статуса и зависимостей в графе.
* Эксперименты с PET на наборах данных Twitter и DBLP показали, что этот подход эффективен и превосходит существующие методы
  1. Современные подходы

1. **Modeling Emerging, Evolving and Fading Topics using Dynamic Soft Orthogonal NMF with Sparse Representation**

**Авторы: Yong Chen, Hui Zhang, Junjie Wu, Xingguang Wang, Rui Liuбand Mengxiang Lina**

В статье рассматривается модель матричной факторизации под названием SONMFSR (Soft Orthogonal NMF with Sparse Representation), которая полностью использует мягкие ортогональные и разреженные ограничения для статического моделирования тем. Путем введения ограничений на возникающие, развивающиеся и исчезающие темы в SONMFSR, авторы легко получают новую DTM под названием SONMFSRd для динамического анализа событий.

**Преимущества:**

* Использование матричной факторизации сильно улучшило результаты метода.
* Данный метод использовался на практике в реальных приложениях.

Эксперименты на двух корпусах демонстрируют превосходство SONMFSRd по сравнению с некоторыми передовыми DTMs как в обнаружении тем, так и в отслеживании. SONMFSRd показывает большой потенциал в реальных приложениях, где популярные темы в Two Sessions 2015 захватываются и отслеживаются динамически для возможных инсайтов.

1. **Model-based Clustering of Short Text Streams**

**Авторы: Jianhua Yin, Daren Chao, Zhongkun Liu**

Статья представляет алгоритм кластеризации коротких текстовых потоков (MStream), который естественным образом может справиться с проблемой эволюции данных, которая делает модель данных недействительной (дрейф концепции) и проблемой разреженности.

**Преимущества**:

* MStream может достичь передовых результатов, обрабатывая поток всего один раз, и может показать еще лучшую производительность, если мы разрешим несколько итераций каждой партии.
* MStreamF, улучшенный алгоритм MStream с правилами забывания, может эффективно удалять устаревшие документы, удаляя кластеры устаревших партий.

1. **Multilevel Event Detection, Storyline Generation, and Summarization for Tweet Streams**

**Авторы: Poonam Goyal, Prerna Kaushik, Pranjal Gupta, Dev Vashisth, Shavak Agarwal, and Navneet Goyal**

Данный подход рассмотрим более подробно, чем предыдущие.

Из названия видно, что метод решает несколько задач: Multilevel Event Detection, Storyline Generation, and Summarization.

1. Multilevel Event Detection

Эта задача решается с помощью онлайн инкрементального алгоритмы кластеризации. Он идентифицирует мелкомасштабные события в виде небольших кластеров. Это делается путем кластеризации твитов и поддержания статистики в структуре данных, называемой вектором кластера твитов.

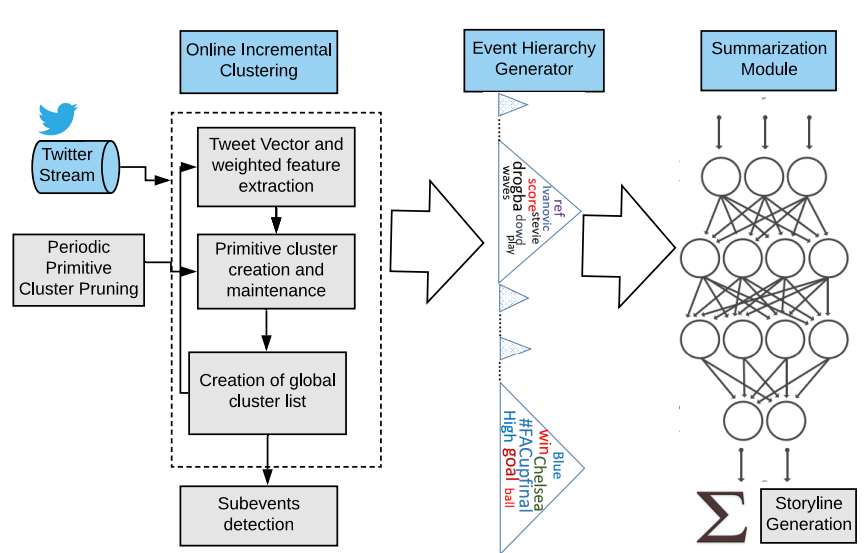
1. Storyline Generation

Здесь используется генератор иерархии событий. Этот модуль генерирует более крупные события в виде иерархий. Это позволяет представить события на разных уровнях абстракции, от самых мелких до самых крупных.

1. Summarization

Модуль резюмирования. Используется модель обучения на основе LSTM для генерации резюме на разных уровнях. Резюме на разных уровнях используются для создания сюжетной линии для события.

Авторы разработали Mythos framework, в котором реализовали выше описанное.

Рисунок 1. Схема работы Mythos framework.

**2. Выявление лучших для использования на практике**

**Кратко резюмируем все современные подходы и сравним их.**

1. [**Model-based Clustering of Short Text Streams**](https://github.com/WHUIR/topic-evolution-on-social-media/blob/master/2020TCSS-Multilevel%20Event%20Detection%2C%20Storyline%20Generation%2C%20and%20Summarization%20for%20Tweet%20Streams.pdf): Этот подход использует модель основанную на кластеризации для обработки коротких текстовых потоков. Он эффективен для группировки коротких текстовых сообщений по темам в реальном времени.
2. [**Modeling Emerging, Evolving and Fading Topics using Dynamic Soft Orthogonal NMF with Sparse Representation**](https://github.com/WHUIR/topic-evolution-on-social-media/blob/master/2020TCSS-Multilevel%20Event%20Detection%2C%20Storyline%20Generation%2C%20and%20Summarization%20for%20Tweet%20Streams.pdf): Этот метод использует динамическую мягкую ортогональную NMF с разреженным представлением для отслеживания тем, которые появляются, развиваются и исчезают. Он эффективен для отслеживания изменения тем во времени.
3. [**Multilevel Event Detection, Storyline Generation, and Summarization for Tweet Streams**](https://github.com/WHUIR/topic-evolution-on-social-media/blob/master/2020TCSS-Multilevel%20Event%20Detection%2C%20Storyline%20Generation%2C%20and%20Summarization%20for%20Tweet%20Streams.pdf): Этот подход, называемый Mythos, предлагает комплексное решение для обнаружения событий, генерации сюжетных линий и создания резюме для потоков твитов. Этот подход эффективен для обнаружения событий и генерации сюжетных линий на основе потоков твитов.

Для программной реализации наиболее подходящий 3 подход, Mythos. Он решает 3 проблемы и даёт гораздо больше результатов по сравнению с другими. Также он является наиболее прогрессивным.

**Заключение**

**Результаты работы**

В данной работе было рассмотрено 7 подходов для эволюции темы, 4 старых и 3 современных. Для каждой выявлены её главные преимущества и недостатки. Сделан сравнительный анализ методов и выявлен лучший среди них - [Multilevel Event Detection, Storyline Generation, and Summarization for Tweet Streams](https://github.com/WHUIR/topic-evolution-on-social-media/blob/master/2020TCSS-Multilevel%20Event%20Detection%2C%20Storyline%20Generation%2C%20and%20Summarization%20for%20Tweet%20Streams.pdf). В дальнейшем планируется реализовать этот метод, а также исследовать новые подходы.

**Список использованных источников**

1. Dynamic Topic Models (2006) David M. Blei и John D. Lafferty
2. Streaming First Story Detection with application to Twitter (2010) Sasa Petrovic, Miles Osborne, Victor Lavrenko
3. PET: A Statistical Model for Popular Events Tracking in Social Communities (2010) Cindy Xide Lin, Bo Zhao, Qiaozhu Mei, Jiawei Han
4. Generating Event Storylines from Microblogs (2012) Chen Lin, Chun Lin, Jingxuan Li
5. Modeling Emerging, Evolving and Fading Topics using Dynamic Soft Orthogonal NMF with Sparse Representation (2015) Yong Chen, Hui Zhang, Junjie Wu, Xingguang Wang, Rui Liu and Mengxiang Lina
6. **Model-based Clustering of Short Text Streams (2018) Jianhua Yin, Daren Chao, Zhongkun Liu**
7. **Multilevel Event Detection, Storyline Generation, and Summarization for Tweet Streams (2020) Poonam Goyal, Prerna Kaushik, Pranjal Gupta, Dev Vashisth, Shavak Agarwal, and Navneet Goyal**