У статті розглянута проблема обробки дублікатів і сюжетних ланцюжків новин при виділенні унікальних подій у новинному потоці. Запропоновано метричні критерії оцінки ступеня близькості новин. Сформульовано алгоритм обробки новинного потоку

Ключові слова: новинний потік, синтаксичні моделі, лексеми, критерій близькості

В статье рассмотрена проблема обработки дубликатов и сюжетных цепочек новостей при выделении уникальных событий в новостном потоке. Предложены метрические критерии оценки степени близости новостей. Сформулирован алгоритм обработки новостного потока

Ключевые слова: новостной поток, синтаксические модели, лексемы, критерий близости

УДК 004.021

ПОДХОД ВЫДЕЛЕНИЯ СОБЫТИЙ В НОВОСТНОМ ПОТОКЕ

И.А. Черенков
Аспирант*
Контактный тел.: (057) 707-64-74
E-mail: igor.cherenkov@gmail.com
С.В. Орехов
Кандидат технических наук, доцент*
Контактный тел: (057) 707-64-74
E-mail: osv@kpi.kharkov.ua
*Кафедра системного анализа
Национальный технический университет
"Харьковский политехнический институт"
ул. Фрунзе, 21, г. Харьков, Украина, 61002

1. Введение

Для успешного функционирования предприятия в рыночной среде необходимо формирование конкурентных преимуществ, которые, в частности, могут быть обеспечены за счёт грамотной ценовой политики. Одной из подзадач ценовой политики является задача ценовой разведки, которая может быть реализована с помощью автоматического прогнозирования значения цены на основе новостного потока [1, 2]. При этом возникает ряд трудностей при обработке новостного потока связанные с выделением уникального события в множестве дублирующих и дополняющих друг друга новостей. Для автоматического выделения дублирующих и сюжетных новостей необходима формулировка соответствующих метрических критериев близости новостей.

2. Цель и задачи исследования

Сформулируем метрические критерии близости двух новостных объектов и алгоритм объединения новостей в кластер.

3. Основная часть

Обработка любого новостного потока, как множества текстовых объектов, предполагает формирование множества словарей лексем, соответствующей тематики, необходимых и достаточных для однозначно выделения событий в новостном потоке [3].

Каждой новости в соответствие ставится уникальная категория, отражающая тип произошедшего события.

Обозначим множество лексем, формирующих словари категорий, $L=\{l_j\}, j=\overline{1,J}$, где l_j - словарь синонимических лексем. Для множества категорий событий K обозначим соответствующие им словари лексем $V^k=\{l_j^k\}, k\in K$.

При этом одни и те же лексемы могут формировать разные словари категорий $V^k \cap V^r \neq 0; k\neq r; r,k\in K$, что делает невозможным однозначное выделение категории новости исключительно на словарях лексем. Необходимая точность идентификации достигается за счёт использование аппарата синтаксических моделей грамматик непосредственных составляющих применительно к анализу названия новости и лида, содержащих всю необходимую информацию о событии. Эффективность такого подхода заключается в том, что для новостей как специфических контейнеров информации в разрезе конкретной отрасли возможно формирование полного множества синтаксических моделей без значительных ресурсных затрат [4, 5]. Составление же некоторого универсального классификатора для любой предметной области не представляется возможным.

Анализ, основанный на синтаксических моделях, позволяет в первичном приближении оценить описываемое в новости событие, однако выделение уникального события возможно только при полной обработке новостного потока. Проблематика обработки новостного потока заключается в том, что большинство данных в новостях носят неметрический характер, а потому необходим переход от неметрического представления данных в новости к метрическому, в котором возможно описание универсального критерия близости двух новостей.

Обработка новости на основе синтаксических моделей предполагает формирование трёх векторов: вектора $\vec{n}$ лексем, отражающего информацию о том: что, где, когда и с кем происходит в новости; вектора $\ddot{\vec{n}}$ восстановленного на основе $\vec{n}$; вектора $\dddot{\vec{n}}$ лексем новости, уточняющие вектор $\vec{n}$.

Вектор $\vec{n}$ всегда имеет фиксированную структуру $\vec{n}=(d,c,G,E,M)$, где d – содержит дату новости (когда), c – тип, категория события (что), G - география

события (где), E - множество контрагентов новости, M - множество товаров (рынков), упомянутых в новости (с кем).

Для объединения двух новостей-дубликатов в одно событие первостепенным условием является совпадение категорий новостей. Дата новостей, отражающих событие должна различаться в пределах порогового значения $d_{treshold}$, отражающего динамику рынка. Степень близости по остальным значениям осуществляться по формуле:

$$F_{primary} = (1+\sum_{i=0}^{I}(m'_j - m''_j)^2)(1+\sum_{i=0}^{I}(e'_j - e''_j)^2)(1+\sum_{i=0}^{I}(g'_j - g''_j)^2) \to \min, (1)$$

где $\vec{M}', \vec{M}''$ - вектора, имеющие в качестве координат $M' \cup M''$ и принимающие значения $m_j = (0,1)$ для тех лексем, что входят в вектора $\vec{n}$ сравниваемых новостей, $\vec{E}', \vec{E}''$ - соответственно вектора контрагентов, $\vec{G}', \vec{G}''$ - вектора географии и рынков новостей, при этом значения географии событий должны быть приведены к общей размерности.

Очевидно, что в большинстве случаев вектора $\dot{\vec{n}}$ будут отличаться и полного совпадения, когда критерий равен единице, наблюдаться не будет. Данная проблема может быть решена либо путём расчета экспертных оценок порогов объединения новостей, либо за счёт формирования порогового значения аналитическим способом.

Предлагается следующий аналитически способ расчёта коэффициента погрешности α, основанный на восстановленном векторе новости $\ddot{\vec{n}}$ прочих лексем. К множеству таких лексем следует отнести те лексемы, что входят в новость, но при этом не отражают само событие непосредственно, а лишь уточняют характер события, в частности, указывают на направленность события во времени, на характер, силу воздействия и т.д.

Для векторов $\ddot{\vec{n}}$ прочих лексем, сформируем множества словарей синонимических лексем, упоминаемых в новостях, $L' \cup L''$, на основе которых сформируем векторы $\vec{L}', \vec{L}''$, такие что $L_i = \begin{cases} 0, l \notin \ddot{\vec{n}} \\ 1, l \in \ddot{\vec{n}} \end{cases}$, откуда, критерий близости векторов принимает вид

$$F_{tertiary} = \sum_{i=0}^{I}(L''_i - L''_i)^2 \to \min . \quad (2)$$

На основании $F_{tertiary}$ может быть получен коэффициент α:

$$\alpha = \frac{\rho(\dddot{\vec{n}}_U)}{F_{tertiary}}, \quad (3)$$

где вектор $\dot{\vec{n}}_U = \dot{\vec{n}}' \cup \dot{\vec{n}}''$, $\rho(\dddot{\vec{n}}_U)$ - длина вектора.

Смысл коэффициента α заключается в том, что чем больше степень сходства по третичному критерию, тем больше вероятность того, что новости описывают одно и то же событие, тем меньшая строгость требуется для сходства по первичному и вторичному критериям.

Однако, применение лишь этих двух критериев для оценки степени близости новостей не достаточно, поскольку довольно часто информация в новости о событии носит неполный характер.

Для большей точности оценки степени близости новостей введём вектор $\ddot{\vec{n}}$ и критерий близости, основанный на нём с целью разрешения проблемы неполноты данных о событии в новости.

Вектор $\ddot{\vec{n}}$ формируется на основе $\dot{\vec{n}}$, исходя из собранной информации о предметной области отрасли, в частности, учитываются какие контрагенты на каких рынках оперируют с каким товаром т.д. Для восстановленных векторов можно рассчитать критерий аналогичный первичному (1):

$$F_{secondary} = (1+\sum_{i=0}^{I}(m'_j - m''_j)^2)(1+\sum_{i=0}^{I}(e'_j - e''_j)^2)(1+\sum_{i=0}^{I}(g'_j - g''_j)^2) \to \min . (4)$$

Откуда, объединяя все три критерия в один (1-4), можно получить метрический критерий близости новостей:

$$F = \begin{cases} c' = c'', \\ |d' - d''| \le d_{treshold}, \\ F_{primary} \le \alpha, \\ F_{secondary} \le \alpha \cdot F_{primary} . \end{cases} \quad (5)$$

Критерий интерпретируется следующим образом: $F_{primary} \le \alpha$ означает, что новости должны быть схожи по первичному вектор тем больше, чем меньше они схожи по третичному критерию; $F_{secondary} \le \alpha \cdot F_{primary}$ означает, что сходство по восстановленному вектору $\ddot{\vec{n}}$ должно быть больше (в связи с большим числом координат), чем для исходного вектора $\dot{\vec{n}}$ в пределах погрешности α.

Алгоритм выделения кластера новостей, описывающих уникальное событие в потоке новостей приведен на рис. 1.

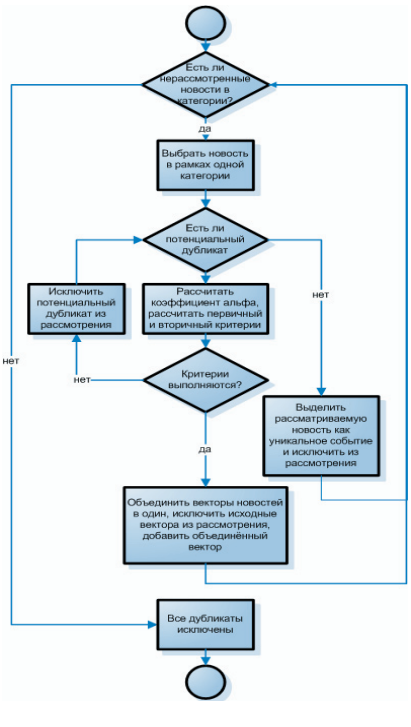

Рис. 1. Алгоритм выделения кластера события

Для каждой новости (рис. 1) формируется множество потенциальных дубликатов из числа новостей входящих в одну и ту же категорию, после чего в рамках одного множества дубликатов новости попарно проверяются на схожесть. В том случае, если две новости не схожи, то сравниваемая новость исключается из списка потенциальных дубликатов, в противном случае происходит объединение векторов двух новостей в новый вектор на основе следующего правила: если две новости описывают одно событие и незначительно отличаются, то объединение векторов этих новостей есть более точное описание события. После рассмотрения одной новости в рамках своего множества дубликатов, в том случае, если были выделен кластер события, все векторы новостей, вошедших в кластер, заменяются на вектор кластера события, что исключает формирование дубликатов событий.

Выделение событий в сюжете новостей будет отличаться от поиска уникального события среди дубликатов, лишь правилом проверки времени поступления новости. Очевидно, что даты должны быть последовательными в пределах той же погрешности $d_{treshold}$. В то же время вопрос выделения сюжетных цепочек в новостном потоке некоторой отрасли при автоматическом прогнозировании является открытым: даже если по причине одного важного события, произошёл ряд менее важных производных событий, то в прогнозировании проще анализировать всё множество событий, а не выделять весомое событие первоисточник. Требуются дальнейшие исследования по обработке новостных сюжетов отраслевых потоков новостей при автоматическом прогнозировании цен.

4. Выводы

Для безошибочного объединения новостей в событийные кластеры необходимо использовать разработанный алгоритм и критерий оценки степени близости.

Высокая эффективность работы алгоритма и критериев возможна за счёт минимизации субъективного фактора: применения синтаксических моделей новостей и расчёта схожести новостей без использования экспертных оценок.

Метрическая природа критериев позволяет их использование в автоматической обработке новостного потока для последующего ценового прогнозирования.

Точность работы подхода в значительной степени зависит от качества первичной обработки новости: качества синтаксических моделей новостей и словарей лексем.

Литература

1. Черенков, И. А. Прогнозирование на основе новостного потока посредством ассоциативных правил [Текст] / Черенков И. А. Общегосударственный научно-производственный журнал. Энергосбережение. Энергетика. Энергоаудит.: Харьков. – 2012. – №11 (105). – С. 38-42.
2. Черенков, И. А. Обоснование прогнозирования цен полимеров посредством новостного потока [Текст] / Черенков И. А., Орехов С.В. Восточно-европейский журнал передовых технологий.- Харьков: 2010. - № 5/7 (47). - С. 18-21.
3. Черенков, И. А. Автоматический поиск данных из новостей на примере рынка полимеров [Текст] / Черенков И. А., Орехов С.В. Системы обработки информации: Харьков. - 2011. - №8. - С. 156-159.
4. Мельчук, И. А. Опыт теории лингвистических моделей смысл-текс. Семантика. Синтаксис [Текст] / Мельчук И.А. - М.: Высш. шк., 1999. - 345 с.
5. Н. Хомский Аспекты теории синтаксиса // Хомский Н. - Изд-во БГК им. И.А. Бодуэн Де Куртенэ. - 1999. - 254 с.

Abstract

Market price forecasting allows effective manage of pricing policy and acquire competitive advantage. Most of existing price forecasting approaches are based either on experts' opinions or on raw price data models. Neither of this approaches allows to get a high forecasting accuracy due to nature of price behaving since price reflects events in real world. Possible solution could be in usage of news based forecasting models. Such forecasting models require processing of news streams.

Processing news streams is a complex task because reflection of event in the news isn't very precise therefore there is a need in development of proper news data processing methods. Main problem in news data processing is filtering news duplicates and plots. One of the possible approaches in news' processing is based on extracting lexemes from news header and first sentence and their further processing.

By forming three vectors based on extracted lexemes for the news it is possible to develop an efficient criteria for duplicates detection. By itself criteria doesn't include any kind of expert opinion for similarity detection except the basic processing logic.

Developed approach allows to extract events data from news stream sufficient for price forecasting

Keywords*: news flow, the syntactic model, tokens, proximity criterion*