Проект "Sign"

**Термины и понятния**

Пространство слов - n-мерное пространство, в котором каждая точка представляет значение какой-либо новости. Координатами должны служить ключевые слова, например, "кризис, стабильность, улучшение, ухудшение" и т.д. Базис пространства определится после первого запуска Парсера и получение реальных данных. Предполагается, что в данном пространстве похожие новости должны располагаться рядом (|**a**-**b**| < e, где **a**, **b** - векторы, || - геометрическое расстояние, e - параметр определяющий похожи данные новости или нет)

В данном пространстве мы строим функцию Амплитуды. Это будет функция многих переменных показывающая какую амплитуду дает данная новость. Учитывая свойство пространства слов, функция должна быть непрерывной и дифференцируемой. Таким образом используя известную амплитуду для нескольких новостей с помощью экстраполяцию мы сможем распространить данную функцию на интересующую нас область пространства.

Влияние новосте - функция времени, позволяющая в каждый момент времени t > t0 вычислить насколько данная новость изменила текущий курс валют. t0 - момент появления новости. Так как мы работаем с дискретными данными, то и функция должна иметь значения только в определенные моменты времени t1, t2, ... tn. Влияние различных новостей в каждый момент времени складывается арифметически.

Активная новость - в момент времени t новость называется активной, если t>t0 и амплитуда влияния больше чем Amin - минимальное влияние новости. Если амплитуда новости меньше некоторого заданного значения Amin, то данной новостью мы можем пренебречь.

**Архитектура программы**

Программа состоит из двух частей - Парсер и Визор.

Архитектура Парсера

Архитектура Визора

**Алгоритм работы**

На вход подается текст новости, на выходе вектор в пространстве "слов". В результате работы парсера мы должны получить некие области "похожих" новостей. Именно эти вектора подаются на вход Визора. Также вместе с новостью мы получаем амплитуду. Амплитуда - это разность цены между следующим баром и текущим. Мы думаем, что новости влияют на цену следующим образом - есть начальная амплитуда А0, далее влияние новости на цену падает приблизительно экспоненциально. И полностью падает через время Т, определяемое экспериментально. Одновременно могут быть активными (влияющими) на цену несколько новостей. В таком случае их амплитуды на текущий момент просто складываются.

В случае, если у нас в один момент времени (один бар)пришло n новостей, мы делаем следующее: обучаем Визор на каждой новости с амплитудой А=А0/n. В случае, если мы попадаем в одну точку пространства новостей (пространство парсера), мы усредняем старое и новое значение.

**Возможные доработки**

При записи амплитуда в пространстве еще один параметр (точность) - надежность новостей в данной области. таким образом, при отображении результата можно будет выводить не только предсказание, но и его достоверность. Если новость встречалась только один раз или не встречалась вовсе - естественно надежность предсказания по данной точке не велика. Однако, попасть в одну и ту же точку пространства тяжело, поэтому при подсчете достоверности используем некий "круг" относительно точки предсказания.

Визор представляет собой нейросеть, задача которой преобразовать вектор новости из пространства парсера в предполагаемую амплитуду для этой новости.

**Цикл работы системы:**

1. Формирование нового бара (i)
2. Проверяем появились ли новые новости на i-1 баре
3. Если да, получаем текст новости, передаем в парсер, парсер передает вектор в Визор. На входе в Визор мы имеем вектор характеризующий новость и амплитуду А0. В результате для новой новости мы находим воздействие новости на цену в виде A = A0\*exp(-t/A0).
4. Проверяем есть ли активные новости в данный момент. Строим продолжение графика цены учитывая только новости.
5. Рассчитываем вероятность (точность) того, что новость правильная.
6. Выводим результаты работы

**Описание модели**

Рассмотрим новости в повседневной жизни. Когда мы только узнаем новую новость она оказывает самое сильное влияние на нас, так как нам необходимо ее обработать, понять и принять решение учитывая новую информацию.

Новости можно разделить по их важности. Важность новости показывает насколько сильно новая информация изменяет текущую ситуацию. Например новость о том, что завтра будет облачно вряд ли сможет испортить планы сходить на прогулку, поэтому новость неважная, однако штормовое предупреждение игнорировать нельзя - это важная новость.

Попробуем перенести понятие важности в нашу модель. Как уже было сказано выше, наиболее сильное воздействие новость оказывает как только она появилась. Переводя на язык форекса - наиболее сильное воздействие новость оказывает на следующий бар, после своего появления. Амплитудой воздействия назовем разность между следующим и текущим барами новости.

Теперь рассмотрим как влияние новости меняется с течением времени. Несмотря на важность новости, со временем она теряет свою актуальность. Причины могут быть в появлении новых новостей, опровергающих или корректирующих первоначальную. Новость может предоставлять прогноз на некоторый промежуток времени, за пределами которого не имеет никакой силы. Независимо от типа новости, с течением времени мы "привыкаем" к информации из новости, принимаем ее в расчет, таким образом новость в любом случае с течением времени теряет свою актуальность.

Экспериментально мы выяснили, что лучше всего уменьшение важности новостей описывается убывающей экспонентой. Амплитуда новости будет коэффициентом перед экспонентой. В зависимости от знака амплитуды новость будет увеличивать текущий курс или уменьшать.

Рассмотрим аргумент экспоненты. Так как мы определили, что лучше всего воздействие новости моделируется убывающей экспонентой, возьмем в качестве аргумента -t. Таким образом:

A = A0\*exp(-t);

В таком случае получается, что независимо от того насколько важна новость, в нашей модели интерес к ней будет пропадать одинаково в каждом случае, однако это не так. Важные новости обсуждаются и передаются гораздо дольше, чем новости незначительные. Таким образом скорость угасания интереса к новости должна быть прямопропоциональна амплитуде новости. Отобразим это в формуле:

А = A0\*exp(-t/|c\*A0|)

где с - коэффициент масштаба (изначально можно пложить равным 1),

t отсчитывается с момента выхода новости.

Благодаря этой поправке время действия не важных новостей будет значительно меньше, чем важных.

**Возможные проблемы**

Рассмотрим подробнее ситуация появления двух и более новостей в течение одного бара. В данном случае в качестве входного параметра вместо А1, А2, ... Аn у нас есть только А0 = А1 + А2 + ... + Аn. В данном случае правильно обучить Визор на каждой новости описанным выше способом не получится.

Мы нашли следующий выход - в случае поступления n новостей во время одного бара каждой новости нужно присвоить амплитуду An = A0/n и обучать обычным способом. Причем если мы попадаем в одну и ту же точку пространства парсера дважды, мы усредняем эти два значения. Это необходимо для рассчета точности определения амплитуды по новости.

Естественно случай, когда амплитуды всех n новостей совпадают очень маловероятен, однако данное упрощение позволяет решить поставленную задачу. Рассмотрим два возможных случая:

Разделим новости на две категории - N1 и N2. Рассмотрим две новости из категорий N1 и N2, пришедшие во время одного бара, причем N1 такая, что она либо не повториться вовсе, либо будет повторяться с очень маленькой частотой. Тогда, благодаря параметру точности, новость N1 станет ненадежной. Значит то, что для N1 мы нашли возможно неправильную амплитуду никак не отразиться на работоспособности программы.

Пусть N2 новость, которая повторяется довольно часто. Тогда, вследствии постоянного усреднения, постоянная состовляющая амплитуды останется неизменной, а все шумы обнулятся. Так как при дальнейшем появлении новости N2 амплитуда будет слабо изменяться, а данная амлитуда будет становиться все более надежной (точной).

Благодаря такому усреднению мы сможем правильно определить амплитуды важных для нас новостей и отбросить неважные новости.

Рассмотрим случай, когда новость не оказала никакого влияния на цену, однако мы рассчитали для нее амплитуду. Если данная новость больше не повториться, она относится к категории N1, а значит правильно обрабатывается в нашей модели. Если же такая новость будет из категории N2, то в результате усреднения амплитуда новости будет ~0, т.к. если она не оказывает постоянного воздействия на курс валют - то амплитуда состоит только из случайных значений. А значит влияние данной новости на курс валют будет A = A0\*exp(-t/|A0|), A0 -> 0, exp(-t/|A0|) -> 0, значит A новости получится тождественно равно нулю и данная новость не будет оказывать никакого влияния на курс валют.