Диаграмма, описывающая постановку задачи

Кузьмин А. А.

11 сентября 2014 г.

Диаграмма

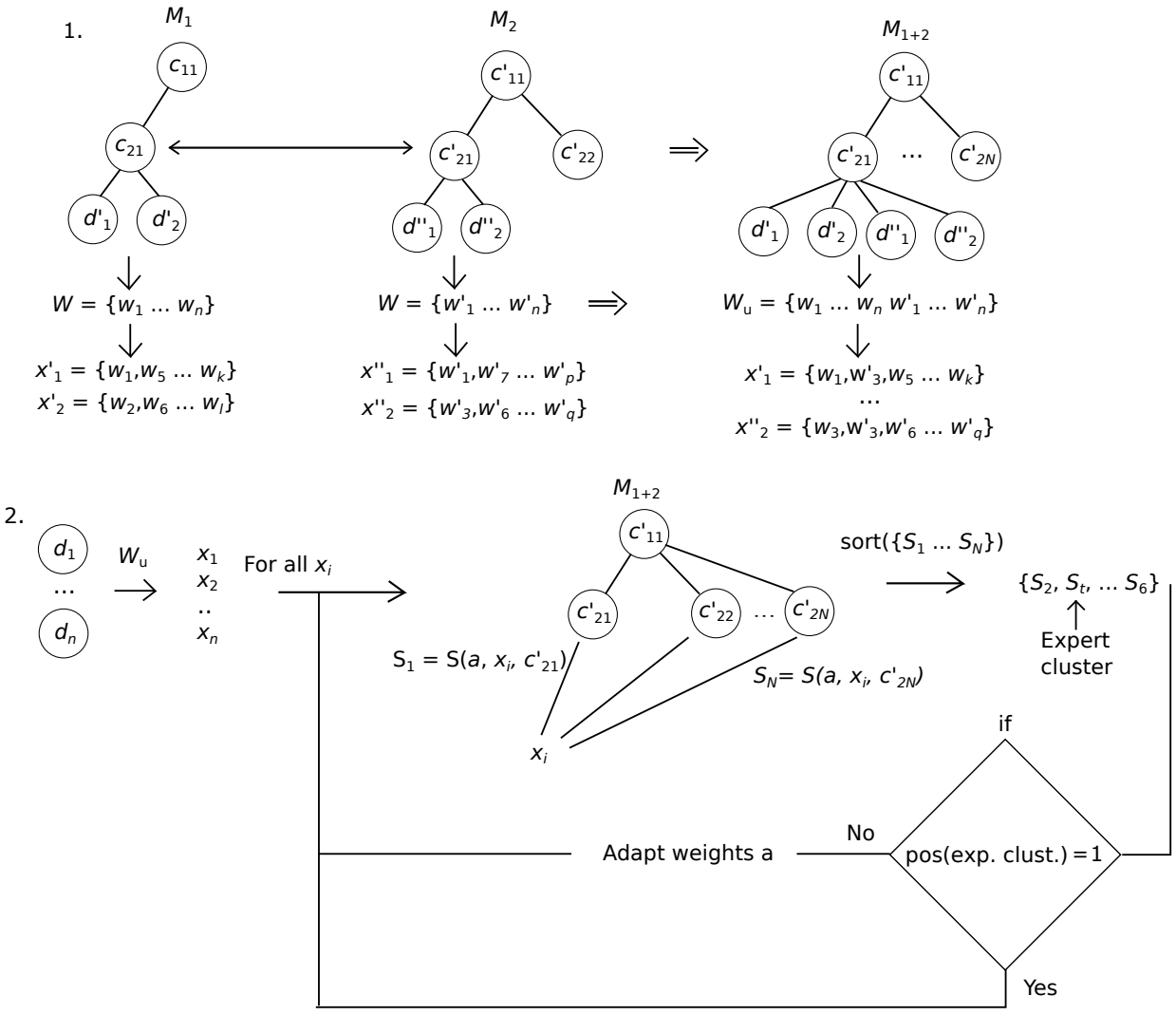

Рис. 1: Схематичное описание задачи

Описание задачи

Имеется несколько экспертных тематических моделей конференции и набор тезисов для следующей конференции. Требуется построить тематическую модель конференции из новых тезисов, схожую с моделями предыдущих конференций.

Конференцию можно схематично представить в виде дерева, в узлах которого стоят кластеры, а листья – документы.

Основные обозначения:

- Коллекция тезисов $D = d_1 \dots d_N$, N – число документов.
- Словарь коллекции $W = w_1 \dots w_n$, n – размер словаря.
- Тематическая модель M.
- Кластеры $c_{i,j}$, где i – уровень кластера, j – порядковый номер на уровне.
- Векторное представление $\mathbf{x}_s$ документа d_s.
- Функция сходства документа и кластера $S(\cdot, \cdot)$.
- Диагональная матрица весов слов из словаря $\mathbf{\Lambda} = diag(\lambda_1 \dots \lambda_n)$

На шаге 1, происходит объединение всех экспертных тематических моделей, кроме одной. При этом, для каждого кластера из модели M_1 ищется соответствующий кластер из M_2. Найденые пары (тройки и т.д.) сливаются в один общий кластер. В случае отсутствия пары для кластера, он переносится в объединенную модель M_{1+2} без изменений. Словари моделей W_i также объединяются.

На втором шаге происходит настройка весов терминов по оставшейся экспертной модели. Для этого все документы представляются ввиде целочисленных векторов, где на месте j стоит число терминов w_j из объединенного словаря W_u в документе. Затем, для каждого документа x_i вычисляется сходство с кластерами объединенной модели. Полученые значения сходства сортируются в порядке убывания. Затем, сравнивается положение экспертного кластера в этой перестановке с 1. Если экспертный кластер имеет наибольшее сходство с данным документом, то система сработала корректно, если же экспертный кластер стоит не на первом месте, то система адаптирует веса признаков, чтобы улучшить дальнейшее качество.