Two Step Method for Grouping News with Similar Topics

Kirill Skorniakov

Orel, 2020

Introduction

Amount of news is rapidly growing up in recent years. People can not handle them effectively. This is the main reason why automatic methods of news stream analysis have become an important part of modern science.

Our work is devoted to the part of the news stream analysis which is called “event detection”.
“Event” is a group of news dedicated to one real-world event

Event Examples

Сюжет 1

* В Цхинвале устраняют последствия шквального ветра ...
* Ураган нанес ущерб жителям Южной Осетии...
* В Цхинвале подсчитывают ущерб, нанесенный стихией ...
*Бибилов: последствия непогоды должны быть устранены в короткое время ...
* Власти Цхинвала восстанавливают поврежденные стихией кровли ...

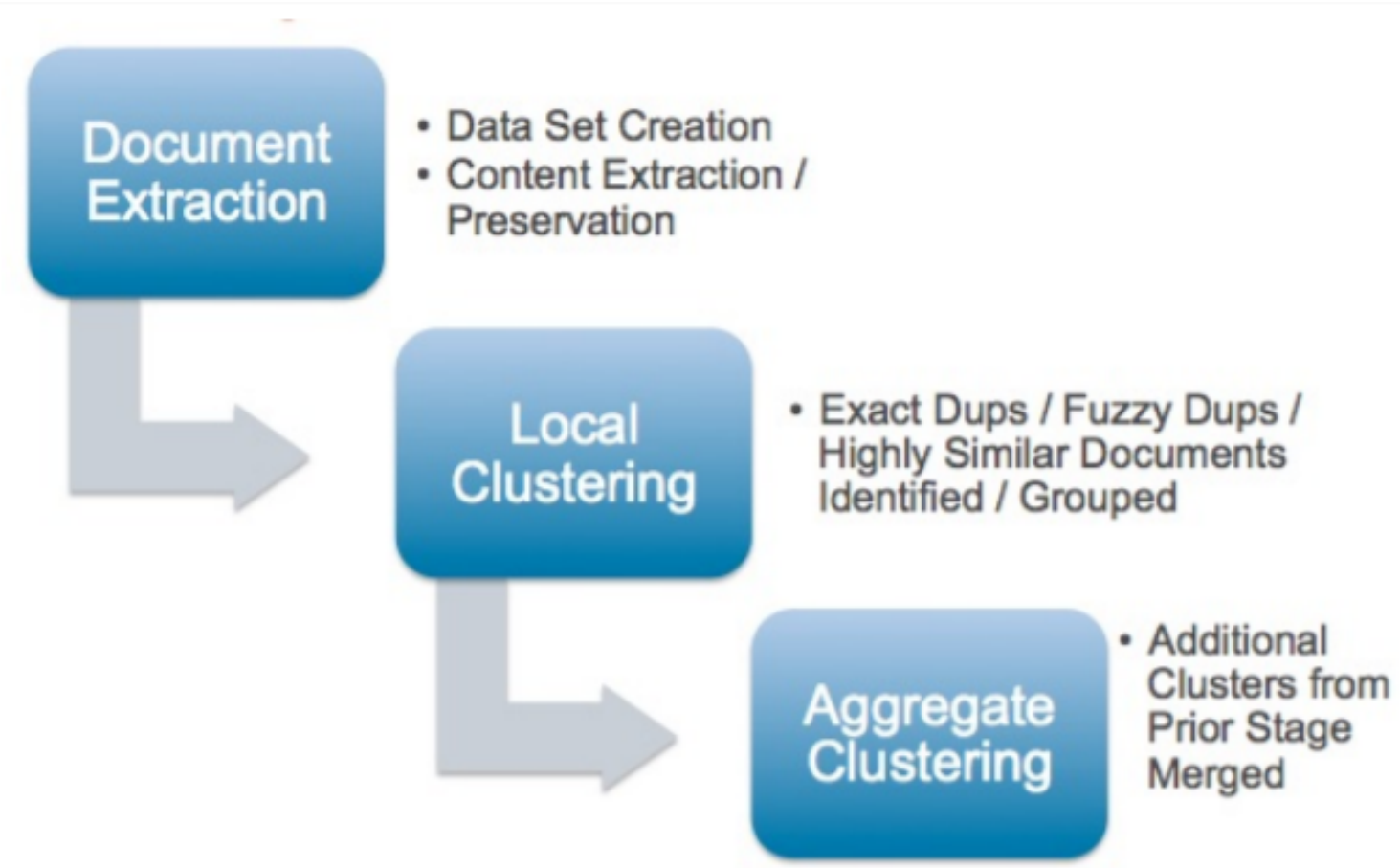
Сюжет 2

* Фестиваль хоровых коллективов прошел в Липецкой области ...
* Крестным ходом и концертом отметят в Хабаровске ...
* День славянской письменности и культуры ...
* День славянской письменности и культуры отметили в Хабаровске ...
* День славянской письменности и культуры отметят в ивановской «научке» ...

Сюжет 3

* В НАТО призвали Россию срочно вернуть Украине моряков и военные корабли ...
* «Оправдания действиям России нет»: НАТО присоединилась к требованиям освободить украинских моряков-провокаторов ...
* Россию призывают незамедлительно освободить украинских моряков ...

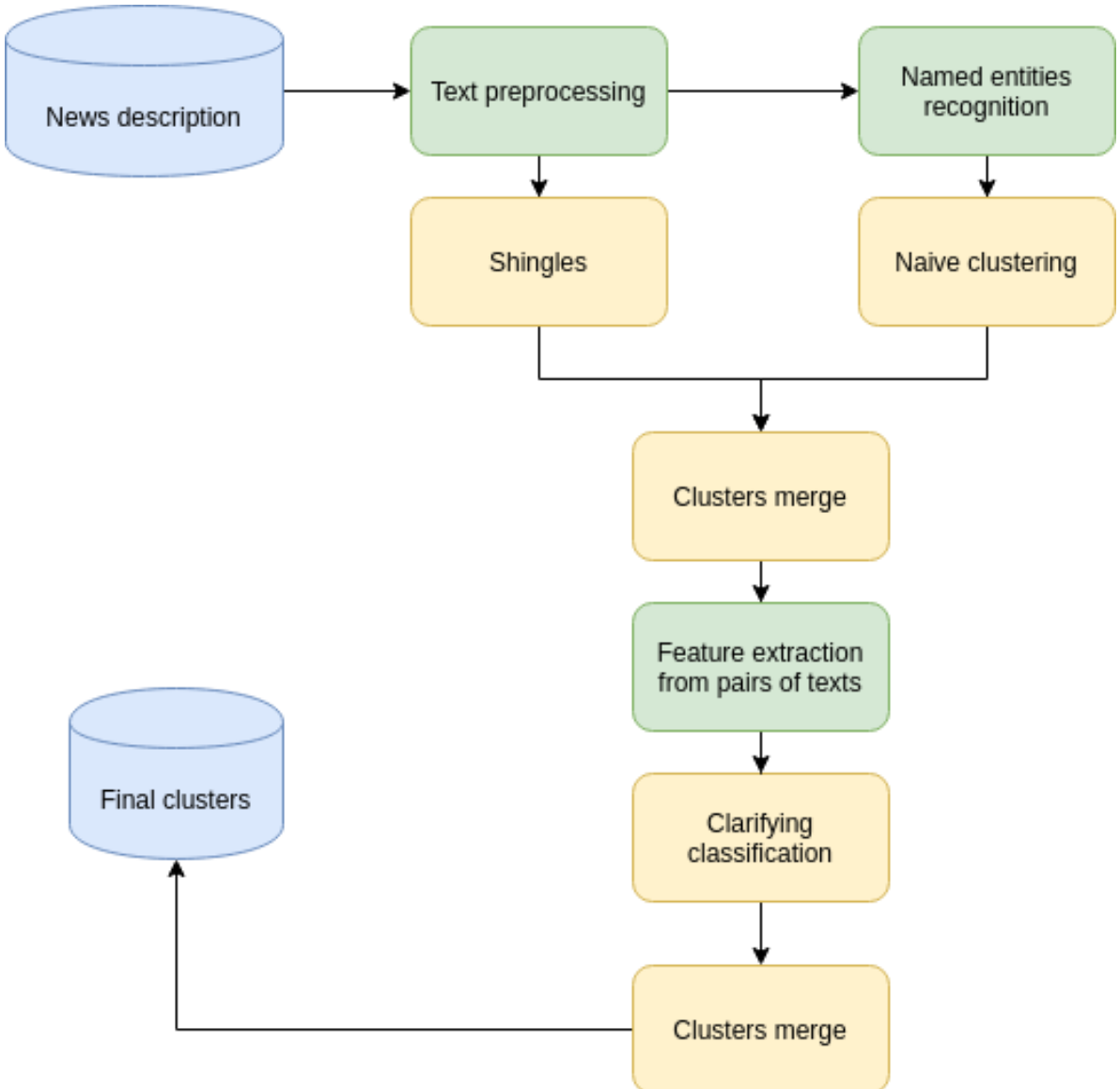
Semi-Supervised Events Clustering in News Retrieval



Problem Statement

- "Event" is a group of news dedicated to one real-world event.
- Only text, title, publications date is available as features.
- Events detects by offline clusterization.

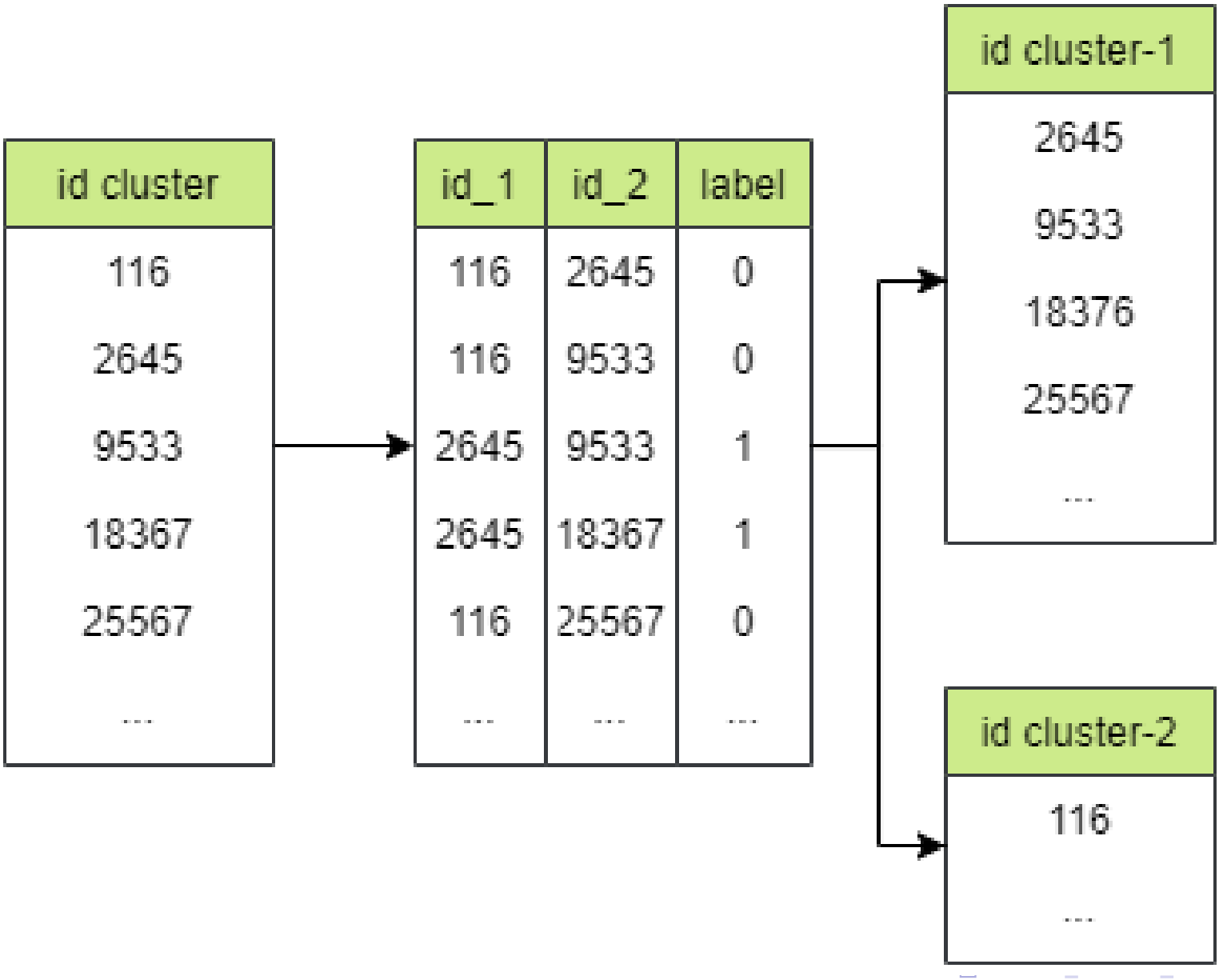
General Scheme



Clusters Merge

1. Create graph of connectivity of texts.
2. Nodes – news, edges – belonging to the same cluster.
3. Compute connected components at this graph.

Clarifying Classification



Dataset

- Crawler of "Yandex News".
- Manual annotation for corrected results.
- Available here[1].

[1] http://talisman.ispras.ru/wp-content/uploads/2020/09/news_events.json_.gz

Input Example

title : "В Москве прошла пресс-конференция, посвященная Чеховскому фестивалю"
url : "https://tvkultura.ru/article/..."
story_url : "https://news.yandex.ru/story/..."
text : "В Москве прошла пресс-конференция, посвященная официальному открытию Международного театрального фестиваля имени Чехова. В первый день работы смотра зрителям представят спектакль Шанхайского центра оперы «Куньцюй» «Пионовая беседка»...."
datetime : 1557878460

Train Set Generation

1. Run simple clustering.
2. Iterate over all pairs at text at each detected cluster.
3. Mark pair as “1” if both texts at the same story at “Yandex News”, “0” otherwise.

Results

SN – Shingles + Naive Clustering
SNC – Shingles + Naive Clustering + Clarifying Classification
SSEC – Semi-Supervised Events Clustering (Conrad and Bender, "Semi-supervised events clustering in news retrieval.")

Method	AMI	Homogeneity	Completeness	V-measure
Shingles	0.45	**0.98**	0.74	0.85
Naive	0.14	0.26	0.79	0.39
SN	0.12	0.22	0.80	0.34
SNC	**0.81**	0.91	**0.89**	**0.90**
SSEC	0.42	**0.98**	0.73	0.84

Conclusion

1. A two-stage scheme was proposed that combines clustering and classification methods on news pairs. It is shown that the second stage of pair classification allows combining various methods and achieving the desired balance between homogeneity and completeness.
2. A data method has been proposed for high-quality training of a classifier for refining classification.
3. Labeled dataset of news event detection based on "Yandex News" service was created.