Large-Scale Cross-Document Coreference Resolution

When was the last time you confused a **fruit** with a **giant tech company**?

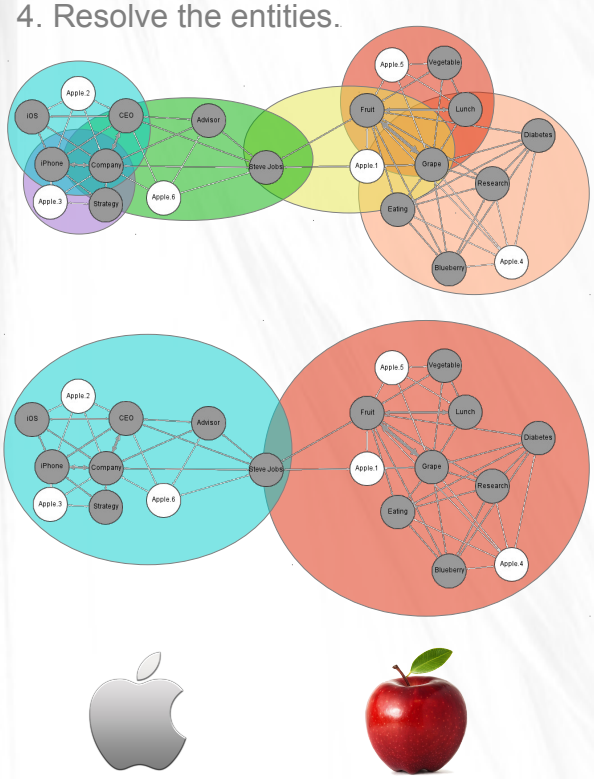
1	Did Steve Jobs really like apple more than other fruits?
2	Apple news: the current iOS for iPhone is going to be improved, said the company's CEO.
3	When Apple releases a new iPhone, the company's current strategy is to …
4	Research showed that eating whole fruits - particularly blueberries, grapes and apples - was 'significantly associated' with a lower risk of type 2 diabetes.
5	Put together a stash of easy, grab-and-go lunch things — cut-up vegetables, grape, fruit that's easy to eat out of hand like apple …
6	Jobs returned to Apple as an advisor, and took control of the company as an interim CEO

Why is it challenging (for a computer)?

- The number of the underlying entities and their identities are unknown.
- The solution space grows exponentially with the number of mentions.
- We have to deal with a multitude of documents with diverse context and different underlying linguistic structures.

Our approach:

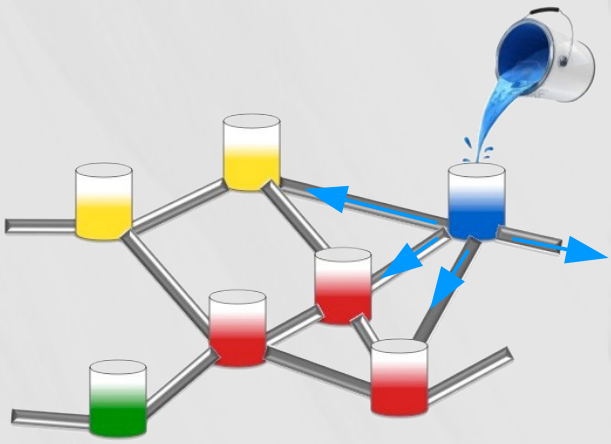
1. Extract the keywords.
2. Construct a graph.
3. Detect communities.
4. Resolve the entities.



Community Detection in Graphs

A diffusion-based approach

1. Initialize the graph with random colors.
2. Diffuse the colors, considering a few policies:
   a. send out less of the dominant color.
   b. send out all the non-dominant colors.
   c. avoid useless flows.
   d. reinforce regions with a dominant color.



Fatemeh Rahimian, Sarunas Girdzijauskas, Seif Haridi

September 2013