Some Subset Selection Problems with Diminishing Returns

Ramakrishna Bairi, Rishabh Iyer, Ganesh Ramakrishnan, Jeff Bilmes

IIT Bombay

15th June, 2015

Subset Selection: Generating Wikipedia Disambiguation Pages

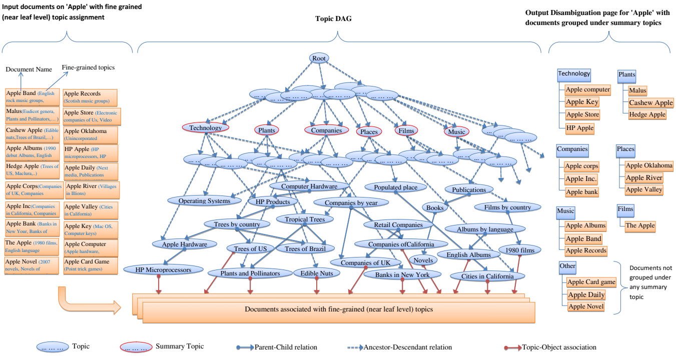

Figure 1: Description of figure on next slide.

Topic Summarization Caption

- On the left, we show many documents related to Apple.
- In the middle, a Wikipedia category hierarchy, shown as a topic DAG, links these documents at the leaf level.
- On the right, we show the output of our summarization process, which creates a set of summary topics (Plants, Technology, Companies, Films, Music and Places in this example) with the input documents classified under them.

Problem Formulation: Basic Notations

- $G(V, E)$: DAG structured topic hierarchy with V topics. E encodes parent-child (*isa*) relationship
- D: Set of documents associated (hard/soft) with one or more of these topics.
- $\Gamma(s)$: Set of documents (transitively) covered by a topic s. Natural extension to set S is $\Gamma(S) = \cup_{s \in S} \Gamma(s)$
- $\Gamma^{\alpha}(s) \subseteq \Gamma(t)$ has path length between a document and s upper bounded by α

Goal

- Given a (ground set) collection V of topics organized in a pre-existing hierarchical DAG structure, and a collection D of documents, chose a size $K \in \mathbb{Z}_+$ representative subset of topics.

Desirable properties

- Goal: Identify summary set of topic $S \subseteq V$ with following properties.
- **Coverage:** S should cover most of the documents. A document d is said to be covered by a topic t if $d \in \Gamma(t)$
- **Diversity:** Summaries should be as diverse as possible, When a document is covered by more than one topic, that document is redundantly covered, e.g., "Finance" and "Banking" would be unlikely members of the same summary.
- Summary qualities also involve "quality" notions, including: **Specificity/Clarity/Relevance/Coherence:** These quality measures help us choose a set of topics that are neither too abstract nor overly specific.

Submodular Functions

- A set function $\mathrm{f}(.)$ is said to be submodular if for any element v and sets $A \subseteq B \subseteq V \setminus \{v\}$, where V represents the ground set of elements, $\mathrm{f}(A \cup \{v\}) - \mathrm{f}(A) \geq \mathrm{f}(B \cup \{v\}) - \mathrm{f}(B)$.
- All our functions are monotone submodular, unless stated otherwise
- A simple greedy algorithm obtains a $1 - \frac{1}{e}$ approximation guarantee for monotone submodular function maximization
- Formally, we solve the following discrete optimization problem:

$$S^* \in \operatorname*{argmax}_{S \subseteq V : |S| \leq K} \sum_i w_i f_i(S) \tag{1}$$

where, f_i are monotone submodular mixture components and $w_i \geq 0$ are the weights associated with those mixture components. Set S^* is the summary topics scored best.

Coverage Functions

- **Weighted Set Cover Function:** Given $S \subseteq V$, $f(S) = \sum_{d \in \Gamma(S)} w_d = w(\Gamma(S))$, assigns weights to the documents based on their relative importance (e.g., in Wikipedia disambiguation, the different documents could be ranked based on their priority)
- **Feature-based Functions:** Represent coverage in feature space. Given $S \subseteq V$ and a set of features U, $m_u(S)$ is the score associated with the set of categories S for feature $u \in U$.
  - U could represent TFIDF features over the documents.

  $f(S) = \sum_{u \in U} \psi(m_u(S))$, where ψ is a concave (e.g., the square root)

Similarity-based Functions

- Defined through a similarity matrix: $\mathbf{S} = \{s_{ij}\}_{i,j \in V}$. Given $i, j \in V$, $s_{ij} = |\Gamma(i) \cap \Gamma(j)|$, (number of documents commonly covered)
- **Facility Location:** $f(S) = \sum_{i \in V} \max_{j \in S} s_{ij}$, is a natural model for k-medoids and exemplar based clustering.
- **Penalty based diversity:** A similarity matrix may be used to express a form of coverage of a set S but penalized with a redundancy term: $f(S) = \sum_{i \in V, j \in S} s_{ij} - \lambda \sum_{i \in S} \sum_{j \in S,} s_{i,j}$; here $\lambda \in [0, 1]$.

Quality Control (QC) Functions

- We define the quality score of the set S as $F_q(S) = \sum_{s \in S} f_q(s)$, where $f_q(s)$ is the quality score of topic s for quality q. Therefore, $F_q(S)$ is a modular function in S.
- **Topic Specificity:** The farther a topic is from the root of the DAG, the more specific it becomes: $f_{\text{specificity}}(s) = s_h$ where s_h is the height of topic s in the DAG.
- **Topic Clarity:** The fraction of descendant topics that cover one or more documents: $f_{\text{clarity}}(s) = \frac{\sum_{t \in \text{descendants}(s)} [\![\Gamma(t) > 0]\!]}{|\text{descendants}(s)|}$, where $[\![.]\!]$ is the indicator function.
- **Topic Relevance:** A topic is considered to be better related to a document if the number of hops needed to reach the document from that topic is lower. Given any set $A \subseteq D$ of document, and any topic $s \in V$: $f_{\text{relevance}}(s|A) = \text{argmin}_\alpha \{\alpha : A \subseteq \Gamma^\alpha(s)\}$.

QC Functions as Barrier Modular Mixtures

- A modular function for every QC function: $f^{\alpha}_{\text{specificity}}(s) = \begin{cases} 1 & \text{if the height of topic } s \text{ is at least } \alpha \\ 0 & \text{otherwise} \end{cases}$ for every possible value of α. This creates a submodular mixture with as many components as the number of possible values of α. In our experiments with Wikipedia, we had α varying from 1 to 120 stepping by 1, adding 120 modular mixture components. Similarly, we define, $f^{\beta}_{\text{clarity}}(s) = \begin{cases} 1 & \text{if the clarity of topic } s \text{ is at least } \beta \\ 0 & \text{otherwise} \end{cases}$ for every possible (discretized to make it countably finite) value of β. And, $f^{\gamma}_{\text{relevance}}(s) = f_{\text{cov}}(s|\Gamma^{\gamma}(s))$, where $f_{\text{cov}}(.)$ is the coverage submodular function and $s|X$ indicates coverage of a topic s over a set of documents X.

Fidelity Functions

- A function representing the fidelity of a set S to another reference set R is one that gets a large value when the set S represents the set R.
- R can be produced from other algorithms such as k-means, LDA and its variants or from a manually tagged corpus.
- **Topic Coherence:** This function scores a set of topics S high when $\Gamma(S)$ resembles the clusters of documents produced by an external source (k-means, LDA or manual). Given an external source that clusters the documents, producing T clusters $L_1, L_2, ..., L_T$ (for T topics), topic coherence is defined as: $f(S) = \sum_{t \in T} \max_{k \in S} w_{k,t}$ where $w_{k,t} = \text{harmonic\_mean}(w^p_{k,t}, w^r_{k,t})$ and $w^p_{k,t} = \frac{|\Gamma(k) \cap L_t|}{|\Gamma(k)|}$ and $w^r_{k,t} = \frac{|\Gamma(k) \cap L_t|}{|L_t|}$. Note that, $w^p_{k,t} \geq 0$ and $w^r_{k,t} \geq 0$ are the precision and recall of the resemblance

Link to Demo

- http://10.129.1.102:
  4020/facets/Pages/Demo/DisambFacetsGen.html