Topic Distillation using Support Vector Data Description

V. Vijaya Saradhi, Harish Karnick and Pabitra Mitra

TRDDC (Pune), IIT Kanpur and IIT Kharagpur

December 16, 2008

Outline

- Introduction
- Topic Distillation Algorithms
  - HITS
  - Bharat & Henzinger Algorithm (B & H)
- Support Vector Data Description (SVDD)
- Experiments and Observations

HITS Algorithm

- Constructs a graph using web pages as vertexes and hyper-links as edges
- Each web page is associated with 'Authority' (sources of information) and 'Hub' (pages with collection of useful links) weights
- Authority and Hub weights are computed iteratively as follows $a_i = \sum_{j \in B(i)} h_j$; $h_i = \sum_{j \in F(i)} a_j$
- Top k weights correspond to good authority and hub pages

B & H Algorithm

- Addresses shortcomings of HITS
  - Equal weights to all the hyper-links
  - Well connected non-relevant web pages (known as *topic drift*)
  - Automatically generated links (also leads to *topic drift*)
- Proposes threshold based heuristic for overcoming the topic drift problem
- 'relevant' pages are obtained in the present work using SVDD

Support Vector Data Description (SVDD)

- Distinguishes one class from rest of the feature space
- Examples are from one class (target class)
- Aims at classifying target examples and non-target examples
- Constructs a hyper-sphere around given data points
- learns 'relevant' page to given query

SVDD Formulation

Primal Formulation:

$$\begin{array}{ll} max_{R,\,O,\,\xi} & R^2 + C\sum_{i=1}^{N} \xi_i \\ subject\ to & \|\varphi(X_j) - O\|^2 \leq R^2 + \xi_j \\ & \xi_j \geq 0, \forall j = 1, \cdots, N \end{array}$$

Dual Formulation:

$$\begin{array}{ll} max_{\alpha} & \sum_{i=1}^{N} \alpha_i \langle \varphi(X_i), \varphi(X_i) \rangle - \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j \langle \varphi(X_i), \varphi(X_j) \rangle \\ subject\ to & \sum_{i=1}^{N} \alpha_i = 1 \\ & 0 \leq \alpha_i \leq C, \forall i = 1, 2, \cdots, N \end{array}$$

- The inner product $\langle \varphi(X_i), \varphi(X_j) \rangle$ is replaced with kernel function $k(X_i, X_j)$

Target Example Identification

- Radius of the hyper-sphere is computed using support vectors $\alpha_i > 0$
- A new data point, say **Z**, is tested by SVDD for acceptance as follows:
  - Computer the distance of **Z** from the origin of the hyper-sphere O
  - If the above distance is less than radius R then **Z** is a target example; else it is not.

Composite Kernel

- The kernel function $k(X_i, X_j)$ is expressed as weighted linear combination of
  - Document similarity kernel
    - $D\,D^T$; where D is a term-document matrix.
  - Co-citation matrix
    - If two documents $X_i, X_j$are cited by ℓ other documents, then the $k(X_i, X_j)$ has a positive score of ℓ

Topic Distillation using SVDD

1. Construct the root set S_R for query q
2. Construct the hyper-sphere using SVDD with p documents from S_R
3. Generate pruned set $S_r \subseteq S_R$using target example identification rule
4. Run HITS on the pruned set $S_r$to obtain top k hubs and authorities

Remarks:

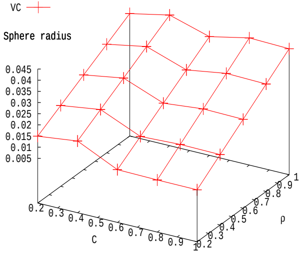
1. Computational over heads
   1.1 Solving the QP using the set S_p
   1.2 Target object identification using the set S_R

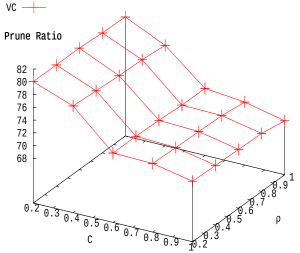
Experimental Results

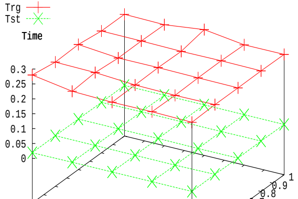
Abbreviation	Query
PA	Parallel Architecture
ZB	Zen Buddhism
VC	Vintage Cars
RE	Recycling Cans
TH	Thailand Tourism

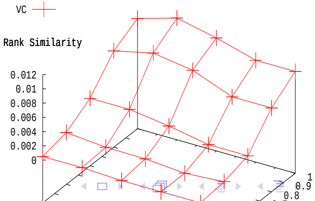
1. SVDD is analyzed along 5 dimensions:
   1.1 Hyper-sphere radius
   1.2 Pruned set S_r
   1.3 Irrelevant pages
   1.4 Computational time
   1.5 Closeness of SVDD to HITS
2. Algorithms Compared: HITS and B & H
   2.1 Precision
   2.2 Relative recall

SVDD Analysis









Irrelevant page example

An example ‘irrelevant’ web page pruned using SVDD.

recycling cans plaese help us to find about cans]

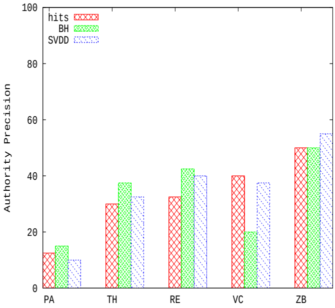
From: (no name) (*no email*)
Date: Mon Mar 04 2002 - 12:02:44 EST

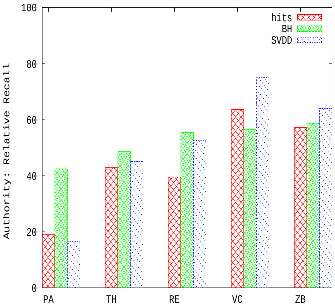
- **Previous message:** ewolfram@infinex.com: "Recycle tip for mesh bags"
- **Messages sorted by:** [date] [thread] [subject] [author] [attachment]

- **Previous message:** ewolfram@infinex.com: "Recycle tip for mesh bags"
- **Messages sorted by:** [date] [thread] [subject] [author] [attachment]

This archive was generated by hypermail 2.1.1 : Mon Mar 04 2002 - 12:02:45 EST

Precision and Relative Recall





- Precision and relative recall figures are obtained through volunteer evaluation
- Shortcomings
  - Number of volunteers is too small
  - No diversity among the volunteers
- SVDD competes with HITS and B&H algorithms

Summary

- Relevant pages from the root set are obtained using SVDD framework
- Authority and hub weights are computed on the pruned set of pages
- Competitiveness of SVDD is experimentally observed
- Computational time in obtaining pruned set is high