SAND2011-5901P

Multi-objective Optimization Approach for Multimodal Information Retrieval

Sandia National Laboratories

Biliana S. Paskaleva, PI, 6923

Arlo Ames, Mentor, 5635

LABORATORY DIRECTED RESEARCH & DEVELOPMENT

Early Career R&D Program

Problem

A principal goal of multimodal information retrieval (MMIR) is extraction of relevant information from large heterogeneous databases.

•A key open problem in MMIR is the development of similarity functions, which maximize the relevance of the ranking with respect to the user's information needs, while minimizing the probability of error.

•Because different similarity functions expose different aspects of the match between the database and queries, we propose to rank the database using a superposition of similarity functions, optimized with respect to utility measures expressing relevance.

Approach

Statement of an abstract MMIR problem

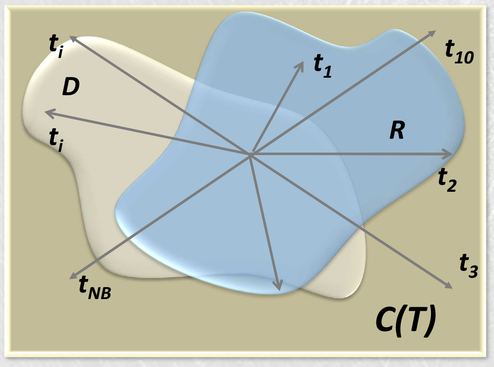
Given: database $D=\{d_1,\dots,d_N\}$ and a query space $Q=\{q_1,\dots,q_R\}$

Elements of *D* and *Q* have *M* different modalities: $d_i=\{d_i^1,\dots,d_i^M\}$, $q_i=\{q_i^1,\dots,q_i^M\}$

$\Sigma=\{S_1,S_2,\cdots,S_k\}$ $S_i:Q\times Q\to[0,1]$ normalized similarity functions

MMIR Problem: Given $q\in Q$, find a set $D(q)=\{d_1,\cdots,d_n\}\subseteq D$, ranked by a superposition of S_i, which maximizes a select utility measure μ.

Generalized Vector Space Model (GVSM)



Token *t* - an elementary piece of information that cannot be broken into smaller tokens.

$T=\{t_1,t_2,\dots,t_k\}$ - token space.

$d=[t_{d1},t_{d2},\cdots,t_{dn}]$ - document = **multiset** of tokens.

d is a proxy of an entity ε.

d(t)- multiplicity of *t* in document *d.*

C(T) - Problem's corpus = **multiset** of documents.

$D=[d_i]_{i=1}^n$, $R=[r_i]_{i=1}^m\subset C(T)$ - databases.

Key elements of the GVSM approach developed in this project

1.T_B* approximates *T: $t\in T_B$ iff there exists $s\in R\cup D \mid t\in s$ We refer to T_B as the basis for *D* & *R*.

2.Generalized indicator function:

$$I:T\times C(T)\to \mathrm{R}^+,\quad I(t,d)=\begin{cases}d(t) \text{ if } t\in d\\ 0 \text{ otherwise}\end{cases}$$

1.Generalized token-to-document indicator matrix

$$\chi=\begin{pmatrix}\chi_{11}&\cdots&\chi_{1n}\\ \vdots&\chi_{ms}&\vdots\\ \chi_{N_{B1}}&\cdots&\chi_{N_{Bn}}\end{pmatrix}\in \mathrm{R}^{N_B\times n}$$, where $\chi_{ij}=I(t_i,d_j)$ $t_i\in T_B$ and $d_j\in D$

4.Token measure: a mapping $w:T\times C(T)\to \mathrm{R}^+$, such that

$$w(t,c)=\begin{cases}\alpha>0 \text{ if } t\in c\\ 0 \text{ if } t\notin c\end{cases}\quad \forall c\in C(T) \text{ and } t\in T$$

5.Document signature relative to T_B: $W(T_B,s)=(w(t_1,s),\cdots,w(t_{N_B},s))$; $t_i\in T_B$

6. Weighted document's norms (relative to a given token measure *w*)

$$\|s\|_{1,w}:=\|W(T_B,s)\|_1=\sum_{t\in T_B}w(t,s);\quad \|s\|_{2,w}:=\|W(T_B,s)\|_2=\sqrt{\sum_{t\in T_B}w(t,s)^2} \text{ for any } s\in D\cup R$$

5.Extension to intersection of documents

$$\|s\cap r\|_{1,w}:=\left\langle\sqrt{W(T_B,s)},\sqrt{W(T_B,r)}\right\rangle=\sum_{t\in T_B}\sqrt{w(t,s)\,w(t,r)};\ \|s\cap r\|_{2,w}:=\langle W(T_B,s),W(T_B,r)\rangle^{\frac{1}{2}}=\sqrt{\sum_{t\in T_B}w(t,s)\,w(t,r)}$$

5. Extension to union of documents

$$\|s\cup r\|_{1,w}:=\|s\|_{1,w}+\|r\|_{1,w}-\|s\cap r\|_{1,w}=\sum_{t\in T_B}\left(w(t,s)+w(t,r)-\sqrt{w(t,s)w(t,r)}\right)$$

$$\|s\cup r\|_{2,w}:=\sqrt{\left(\|s\|_{2,w}^2+\|r\|_{2,w}^2-\|s\cap r\|_{2,w}^2\right)}=\sqrt{\left(\sum_{t\in T_B}\left(w(t,s)^2+w(t,r)^2-\sqrt{w(t,s)w(t,r)}\right)\right)}$$

Results

Key accomplishments

1.Formulated an abstract definition of the entity resolution problem

Def 1: Given a database D and entity ε, the set of all documents $d\in D$ that are proxies of ε is called the equivalence class of ε in D, D_ε

Def 2: Given databases D and R, for every $r\in R$ find all $d\in D$ which belong to the same equivalence class as r.

2.Developed consistent extensions of set-based similarity measures to documents comprised of token multisets and using norms beyond the L1 norm

✓ *Extension of Jaccard similarity measure* $$J_k(s,r):=\frac{\|s\cap r\|_{k,w}}{\|s\cup r\|_{k,w}},\ k=1,2$$

✓ *Extension of NWI similarity measure* $$N_k(s,r):=\frac{\|s\mid r\|_{k,w}}{\max\{\|s\|_{k,w};\|r\|_{k,w}\}},\ k=1,2$$

✓ *Extension of Dice similarity measure* $$D_k(s,r):=\frac{2\|s\cap r\|_{k,w}}{\|s\|_{k,w}+\|r\|_{k,w}},\ k=1,2$$

Example $J_1(s,r):=\dfrac{\sum_{t\in T_B}\sqrt{w(t,s)w(t,r)}}{\sum_{t\in T_B}\left(w(t,s)+w(t,r)-\sqrt{w(t,s)w(t,r)}\right)}$; for $w(t)$ $J_1(s,r)=\dfrac{\sum_{t\in s\cap r}w(t)}{\sum_{t\in s\cup r}w(t)}$

3.Formulated and implemented a constrained optimization approach to find an optimal superposition of the generalized similarity measures provided by the GVSM framework. Applied the approach to a benchmark entity resolution problem using the e-commerce databases Abt-Buy.com, Univ. Leipzig

http://dbs.unileipzig.de/en/research/projects/object_matching/fever/benchmark_datasets_for_entity_resolution

3.Performed preliminary studies solving the optimization problem for *p=1,2,10,∞.*

$$\min_x\|\mathbf{Cx}-\mathbf{d}\|_p \text{ subject to } \sum_i x_i\le 1 \text{ and } 0\le x_i\le 1$$

$$\mathbf{C}=\begin{pmatrix}\mathbf{C}^1\\ \mathbf{C}^2\end{pmatrix},\quad \mathbf{C}^k=\begin{pmatrix}c_{11}^k&\cdots&c_{1N}^k\\ \cdots&\cdots&\cdots\\ c_{M1}^k&\cdots&c_{MN}^k\end{pmatrix}\to\begin{cases}c_{ij}^1=sim_i(d_j,q_j);\ j=1,\cdots,M\\ c_{ij}^2=sim_i(d_1,q_j);\ j=2,\cdots,M\end{cases},\quad \mathbf{d}=\begin{pmatrix}\mathbf{d}^1\\ \mathbf{d}^2\end{pmatrix}\to\begin{cases}\mathbf{d}_i^1=1,i=1,\dots,M\\ \mathbf{d}_i^2=0,i=1,\dots,M-1\end{cases}$$

Metric/Error	Error Training [%], 100 samp	Error Testing [%], 122 samp
Jaccard-L1 (tfidf)	15	16
Dice-L1 (tfidf)	15	16
NWI-L1 (tfidf)	14	17
Jaccard-L2 (tfidf)	17	14
Dice- L 2(tfidf)	15	14
NWI-L2 (tfidf)	15	14
Jaccard-L2 (idf)	16	14
Dice-L1 (idf)	15	15
NWI-L1 (idf)	17	15
Jaccard-L2 (idf)	21	16
Dice- L2 (idf)	14	14
NWI-L2 (idf)	15	16
Euclidean (tfidf)	35	41
Citiblock (tfidf)	17	18
Minkowski (tfidf)	46	53
Chebychev (tfidf)	38	55
Hamming (tfidf)	16	12

Norm/ Error	Weights	Error Train [%]	Error Test [%]
L1	0 0 0 0 1 0 0 0 0 0	15	14
L2	0 0 0 0 0.99 0 0 0 0.023 0	15	14
L10	0 0 0 0 0.29 0 0 0.14 0.57 0	18	17
L∞	0 0 0 0 0.23 0 0 0 0.22 0.54	25	29.5

Norm/ Error	Weights	Error Train [%]	Error Test [%]
L1	0 1 0 0	17	18
L2	0 1 0 0	17	18
L10	0 0.51 0.48 0	22	25
L∞	1 0 0 0	35	41

Significance

- Lifting of MMIR into an abstract setting separates our approach from existing methodologies, often dominated by ad-hoc or problem-specific algebraic techniques.
- We aim to discover basic design principles adaptable to diverse MMIR problems, arising in strategic NNSA and DOE themes.
  - ✓ **detection of nuclear weapons proliferation,**
  - ✓ **rapid determination of anomalies and bio-threats,**
  - ✓ **effective early detection of physical and cyber attack against energy grids**
- Project success will provide Sandia with MMIR capabilities supporting its national security mission, and also, enable future growth in this area.

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000."

SAND 2010-xxxP

Sandia National Laboratories