
Beyond Bags of Words:

Modeling Implicit User Preferences in Information Retrieval

Donald Metzler and W. Bruce Croft
Center for Intelligent Information Retrieval
University of Massachusetts Amherst



Motivation



Are either of these documents relevant to the query *pet therapy*?

363 the	284 and
260 of	232 to
174 pet	151 in
117 for	88 a
82 that	64 is
59 with	44 by
39 are	37 or
36 from	35 <i>disease</i>
35 have	35 has
35 been	33 new
33 <i>fda</i>	32 as
31 be	30 <i>medicare</i>
29 imaging	29 <i>cti</i>
28 this	27 we
24 reimbursement	23 coverage
23 on	23 was
23 therapy	23 not
22 carcinoma	22 scans

87 the	64 to
54 and	50 of
38 a	26 shelter
26 <i>animal</i>	23 seattle
22 in	19 for
16 are	16 your
15 our	15 i
14 pet	13 <i>animals</i>
13 my	13 bequest
11 or	11 you
11 they	10 new
9 people	9 york
9 <i>dogs</i>	9 therapy
9 be	8 on
8 their	8 like
8 that	8 we
8 all	8 estate
8 who	7 is

More information...



- The URL of document A is:
<http://www.science.doe.gov/bes/Senate/douglas.pdf>
- In document B, there is a picture of a dog
- In document A, pet always appears as PET (i.e. as an acronym)
- In document B, the exact phrase “*pet therapy*” occurs 7 times
- In document A, *pet* and *therapy* never occur together in the same sentence

Full Representations



Senate Committee on Commerce, Science, and Transportation
 Subcommittee on Science, Technology, and Space
 "Emerging Technologies in the New Millenium"

May 12, 1999

Introduction

Chairman Frist and other distinguished Members of the Subcommittee, thank you for this opportunity to testify at this important hearing. My name is Terry Douglass, and I am president of CTI, Incorporated whose home offices are located on Innovation Drive along the Tennessee Technology Corridor between Knoxville and Oak Ridge, Tennessee. Knoxville is the home of The University of Tennessee, and Oak Ridge is the home of the Oak Ridge National Laboratory which makes the location of our headquarters midway between the two institutions significant because of the highly technical nature of our business. CTI is the worldwide commercial leader in providing products and services to the positron emission tomography (commonly referred to as PET) market. PET is a medical diagnostic imaging technology that provides unique and useful metabolic, biochemical, and functional information necessary to the diagnosis and treatment of patients with cancer, heart, and brain disorders.

PET THERAPY

A Dog's Perspective by Curtis, as told to us

My name is Curtis, I'm a proud graduate of the Seattle Animal Shelter. I was lucky enough to be adopted in December 1998. In May of 2000, after we completed obedience school, my mom and I joined the pet therapy team. We visit long-term care facilities in the Seattle area, we dogs bring our parents, who have all been to the shelter. Whimsie, Oriantiana, Virginia Dalton, Animal Care Supervisor, takes us for transportation to make sure we are suitable for the program. I was nervous about the test, but I passed with flying colors!



We make our Pet Therapy visits a few times a month. It's called Pet Therapy because of the awesome benefits animals have on people. And they're just now beginning to see this? Duh! We lower heart rates and stress, and even can improve mobility and motivation in sick and injured people. Personally, I think we therapy dogs are pretty lucky. I get to visit with all these great people who give me love! They tag on my ears, pet my nose, and occasionally I even get a belly rub. Sometimes they have trouble left over in their leg from lunch and they don't mind a relaxing on a few. The patients definitely look forward to our visits.

While they are petting and snuggling me, they often tell my mom about dogs they grew up with or what they used to do for a living. One of my human buddies used to be an animal control officer. Others are retired doctors, ballroom dancers, and more. I like hearing their stories, especially while cleaning up their crumbs and getting ear rubs.

My fellow therapy dogs come in all shapes and sizes. There are several breeds like me and lots of pumbrud dogs, too. My canine friend Tigger does fancy tricks like balancing a treat on his nose for several minutes. The Pet Therapy team here a BL, Bernese, a Great Dane, Cocker Spaniel, a miniature poodle, a Border Collie, a Bernese Mountain Dog, and more! The staffers like seeing all the animals, and I like being able to visit with my canine and human friends.

My job is rewarding, mostly in seeing the happiness and love when they feel my soft fur. And did I mention that I get to nibble the crumbs off their laps? If you are interested in joining us on the volunteer Pet Therapy team, just come to the next shelter orientation and find out how you can be a pet therapy partner!

WISH LIST

The shelter is in need of the following items for our animals. Any and all donations are gratefully accepted.

- Airline Animal Crates
- Bird Toys
- Blankets- towel or quartered (no down, latex, plastic)
- Catnip
- Cat Toys
- Citronella Bark Collar
- Couch or good chair
- Fishy and/or Tuna Cat Food
- Guinea Pig Food
- Hot Water Bottle
- New Age or Classical CD
- Nylon Slip/Onion Collars (Kevlar Collar)
- Rabbit Food
- Salt Licks for Rabbits
- Towels
- Wash Cloths

THANK YOU!

Acknowledgments

We at the Seattle Animal Shelter continue to be impressed by the generosity and compassion of Seattle citizens who care about the health and welfare of our animals. We'd like to say thanks...

Donations

...to everyone who dropped off pet food, blankets, toys and other supplies for the the New York Humane Society and the pets of New York City after the September 11 tragedy.

...to Bruce Small for contributing \$500 to help pay for behavioral modification training for a dog that had been removed from a neglectful home and needed to learn socialization skills.

...to Lilly Chin who donated her expert sewing skills to alter our kennel officers' uniforms.

...to Karin Holmich and Noel Yates of Art Partners for designing and painting our Pet Cam Run and small critter adoption area.

Document Representations



- ◆ Bag of words
 - Most common representation in information retrieval
 - Used by majority of TREC systems
- ◆ Richer features / representations
 - *N*-grams
 - Images
 - PageRank
 - Anchor text
 - Genre
 - Reading level
 - Document structure / formatting

Information Needs



Query Representations



- ◆ Implicit

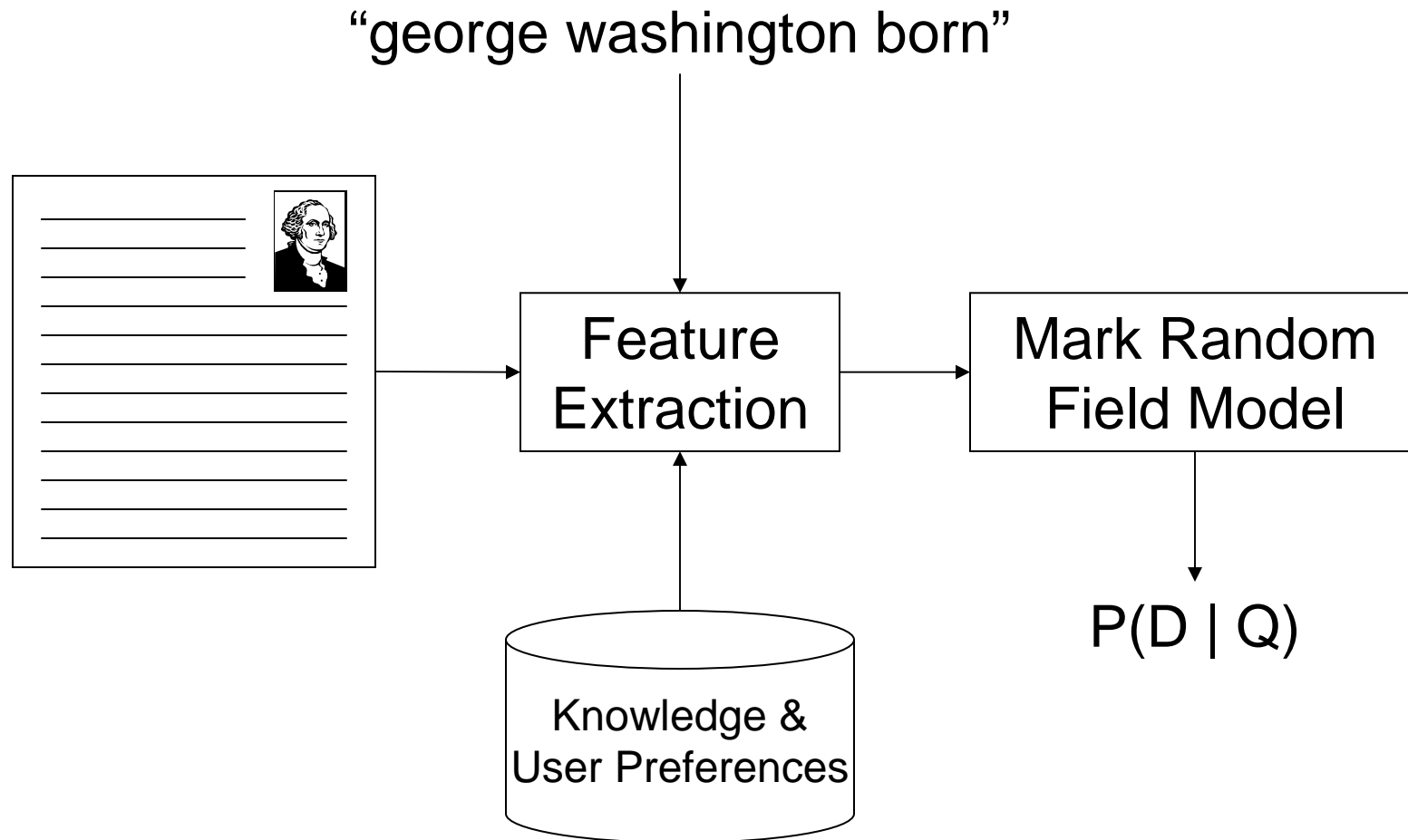
- “george washington born”

- ◆ Explicit

- `#weight[sentence](2.0 #uw8(george washington)
1.0 born
1.0 #any:location)`

- “I want to find **sentences** that contain the terms **George** and **Washington**, in any order, within **8 words** of each other (weighted 2), the term **born** (weighted 1), and **text indicative of a location** (weighted 1).”

Our Model



Our Model



- ◆ Attempts to “reverse engineer” the process that users go through when translating from information need to query
- ◆ Moves away from simple representations
- ◆ Tries to capture implicit user knowledge and preferences via rich features
- ◆ Statistical
 - Formally motivated
 - Supervised parameter estimation

Case Study: Term Proximity



- Model the implicit preference that user's prefer query terms to appear coherently within documents
- Use a simple term proximity model to explicitly extract concepts from query

Query: teaching disabled children

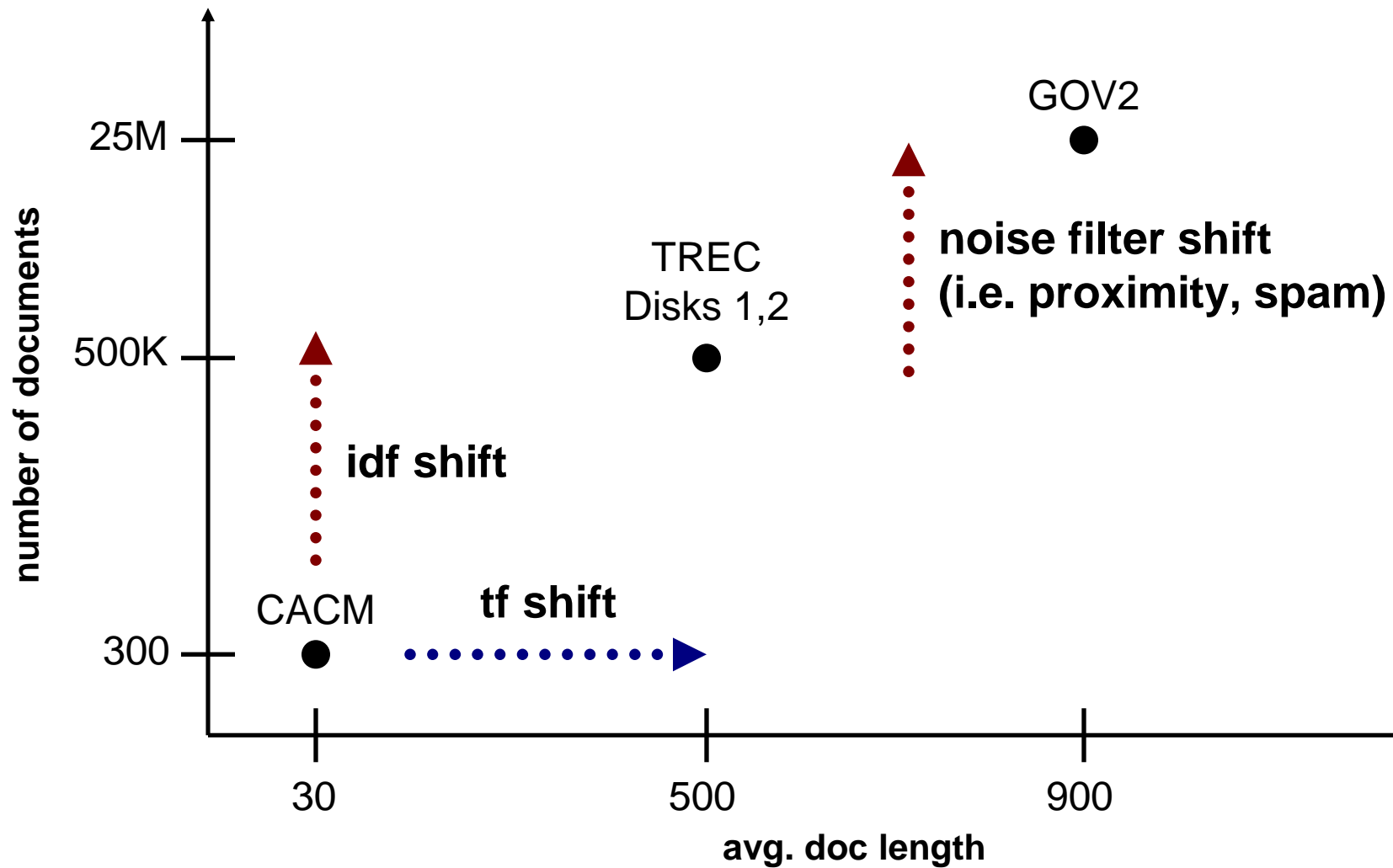
teaching	#uw8(teaching children)
disabled	#uw8(disabled children)
children	#uw8(teaching disabled)
#1(disabled children)	#uw12(teaching disabled children)
#1(teaching disabled)	#1(teaching disabled children)

Results



- ◆ Significant improvements in effectiveness over bag of words and n -gram models
 - Provides insight into why n -gram features for IR and text classification typically do not help
- ◆ Larger improvements seen for larger collections
 - *tf.idf*-based weighting schemes give too much weight to random occurrences of query terms

IR Paradigm Shifts



How AI Can Improve IR



- ◆ Need knowledge rich document and query representations and features
 - Common sense knowledge
 - Better semantic understanding
 - Use of *all* information available (images, hyperlinks, user knowledge/experiences, etc.)
- ◆ Need a better understanding of implicit vs. explicit query representations
 - Users 'trained' to use short keyword queries
 - Can complex queries help?
- ◆ Need to go beyond the text

Conclusions



- ◆ Bag of words models are inadequate for highly effective information retrieval
- ◆ Case study showed that simple term proximity model can significantly improve effectiveness over bag of words models
- ◆ As the amount of information grows, there is a growing need for better modeling of user's implicit preferences *or* a better way of explicitly representing information needs