# Learning Ranking Functions with SVMs

CS4780/5780 – Machine Learning Fall 2013

> Thorsten Joachims Cornell University

T. Joachims, Optimizing Search Engines Using Clickthrough Data, Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD), ACM, 2002. <u>http://www.cs.cornell.edu/People/tj/publications/joachims\_02c.pdf</u>

### **Final Course Projects**

- Now
  - Start thinking of project ideas, anything relevant to the course goes
  - Start recruiting team members
- Oct 22
  - Submit project proposal as group of 3-4 students
- Oct 24
  - Submit peer feedback for proposals
- Nov 21
  - Submit status report
- Dec 5
  - Project poster presentations (evening)
- Dec 11
  - Submit final project report
- Dec 18
  - Submit peer reviews of reports

### Adaptive Search Engines

- Traditional Search Engines
  - One-size-fits-all
  - Hand-tuned retrieval function
- Hypothesis
  - Different users need different retrieval functions
  - Different collections need different retrieval functions
- Machine Learning
  - Learn improved retrieval functions
  - User Feedback as training data



#### Overview

- How can we get training data for learning improved retrieval functions?
  - Explicit vs. implicit feedback
  - Absolute vs. relative feedback
  - User study with eye-tracking and relevance judgments
- What learning algorithms can use this training data?
  - Ranking Support Vector Machine
  - User study with meta-search engine

### Sources of Feedback

- Explicit Feedback
  - Overhead for user
  - Only few users give feedback
    - => not representative

#### Implicit Feedback

- Queries, clicks, time, mousing, scrolling, etc.
- No Overhead
- More difficult to interpret

🛎 Google Search: svm - Microsoft Internet Explorer	<u>- 🗆 ×</u>			
File Edit View Favorites Tools Help	<b>1</b>			
↔ Back ▼ → ▼ 🙆 🖉 🚰 🔍 Search 🗟 Favorites 🛞 Media 🧭 🗟 ▼ 🖨 🔍 ▼ 🗐				
Address 🙆 http://www.google.com/search?sourceid=navclient&ie=UTF-8&oe=UTF-8&q=svm 💽 🔗 G	> Links »			
Google▼ svm 💽 🎆 Search Web ▼ 🎕 Search Site   🐲 PageBank 🕄 ▼ 🗗 65 blocked	t "팀Auto »			
Google Search Directory Nous	<u> </u>			
Searched the web for svm. Results 1 - 10 of about 329,000. Search took 0.29 se	conds.			
Categories: <u>Computers &gt; Artificial Intelligence &gt; Machine Learning</u> <u>Computers &gt; Artificial Intelligence &gt; Neural Networks &gt; Software</u>				
Show stock quotes for Sponsored Links SVM (ServiceMaster Company The)				
Bienvenue sur svm vnunet.fr   - [ Translate this page ] Les forums de SVM. Participez aux grands débats de la rédaction. De vous à vous. Les meilleures réponses sélectionnées sur le forum de SVM svm.vnunet.fr/ - 39k - Mar 1, 2004 - <u>Cached</u> - <u>Similar pages</u>	Today!			
SVM-Light Support Vector Machine SVM-Light Support Vector Machine Hier finden Sie Informationen zu den folgenden Themen: Thorsten Joachims, SVM-light, SVM light, SVMlight, Support Vector Description: Training software for large-scale SVMs. [Free for non-commercial use] Category: <u>Computers &gt; Artificial Intelligence &gt; &gt; Software</u> symlight, joachims, org/ - 3k - Mar 1, 2004 - Cached - Similar pages				
Support Vector Machine Support Vector Machine Older versions are still available from here www-ai.cs.uni-dortmund.de/SOFTWARE/SVM_LIGHT/svm_light.html - 6k - <u>Cached</u> - <u>Similar pages</u>				
ServiceMaster We Are Home ServiceMaster Issues Information on Tax Treatment of Dividends. ServiceMaster Reports 2003 Fourth Quarter Revenues and Profits. ServiceMaster				
www.svm.com/ - 13k - Mar 1, 2004 - <u>Cached</u> - <u>Similar pages</u>				
Kernel Machines Description: A central source of information on kernel based methods, including support vector machines, Gaussian Category: <u>Computers &gt; Artificial Intelligence &gt; Support Vector Machines</u> www.kernel-machines.org/ - 1k - <u>Cached</u> - <u>Similar pages</u>				
SVM Application List SVM Application List. This list of Support Vector Machine applications grows thanks to visitors like vou who ADD new entries svm learning	<b>_</b>			

## Feedback from Clickthrough Data

**Relative Feedback:** Clicks reflect preference between observed links. **Absolute Feedback:** The clicked links are relevant to the query.

(3 < 2),(7 < 2), (7 < 4), (7 < 5), (7 < 6)

- Kernel Machines http://svm.first.gmd.de/
   Support Vector Machine
- http://jbolivar.freeservers.com/
- 3. SVM-Light Support Vector Machine http://ais.gmd.de/~thorsten/svm light/
- 4. An Introduction to Support Vector Machines http://www.support-vector.net/
- 5. Support Vector Machine and Kernel ... References http://svm.research.bell-labs.com/SVMrefs.html
- 6. Archives of SUPPORT-VECTOR-MACHINES ... http://www.jiscmail.ac.uk/lists/SUPPORT...
- 7. Lucent Technologies: SVM demo applet *http://svm*.research.bell-labs.com/SVT/SVMsvt.html
- 8. Royal Holloway Support Vector Machine http://svm.dcs.rhbnc.ac.uk

Rel(1), NotRel(2), Rel(3), NotRel(4), NotRel(5), NotRel(6), Rel(7)

# User Study: Eye-Tracking and Relevance

#### Scenario

- WWW search
- Google search engine
- Subjects were not restricted
- Answer 10 questions
- Eye-Tracking
  - Record the sequence of eye movements
  - Analyze how users scan the results page of Google
- Relevance Judgments

Who discovered the first modern antibiotic? Find the homepage of Emeril - the chef who has a TV cooking program. What actor starred as the main character in the original 'Time Machine' movie? Find the page displaying the routemap for Greyhound buses You are excited to cast your vote in the democratic presidential primary - when can you do so in NY? Find the homepage of Michael Jordan, the statistician. Where is the tallest mountain in NY located? Find the homepage for graduate housing at Carnegie Mellon University. A friend told you that Mr. Cornell used to live close to campus - between University and Stewart Aves does anyone live in his house now; if so, who?

Find the homepage of the 1,000 Acres Dude Ranch.

- Ask relevance judges to explicitly judge the relevance of all pages encountered
- Compare implicit feedback from clicks to explicit judgments

### What is Eye-Tracking?

#### Eye tracking device





Device to detect and record where and what people look at

- Fixations: ~200-300ms;
   information is acquired
- Saccades: extremely rapid movements between fixations

#### Pupil dilation: size of pupil indicates interest, arousal

"Scanpath" output depicts pattern of movement throughout screen. Black markers represent fixations.

#### How Many Links do Users View?



Mean: 3.07 Median/Mode: 2.00

# In Which Order are the Results Viewed?



=> Users tend to read the results in order

### Looking vs. Clicking



=> Users view links one and two more thoroughly / often
=> Users click most frequently on link one

#### Do Users Look Below the Clicked Link?



=> Users typically do not look at links below before they click (except maybe the next link)

#### How do Clicks Relate to Relevance?

- Experiment (Phase II)
  - Additional 16 subjects
  - Manually judged relevance
    - Abstract
    - Page
- Manipulated Rankings
  - Normal: Google's ordering
  - Swapped: Top Two Swapped
  - Reversed: Ranking reversed
- Experiment Setup
  - Same as Phase I
  - Manipulations not detectable

#### . Kernel Machines http://www.kernel-machines.org.

- 2. Support Vector Machine *http://jbolivar.freeservers.com/*
- SVM-Light Support Vector Machine http://ais.gmd.de/~thorsten/svm light/
- 4. An Introduction to SVMs *http://www.support-vector.net/*
- 5. Support Vector Machine and ... http://svm.bell-labs.com/SVMrefs.html
- 6. Archives of SUPPORT-VECTOR... http://www.jisc.ac.uk/lists/SUPPORT...
- 7. Lucent Technologies: SVM demo applet http://svm.bell-labs.com/SVMsvt.html
- 8. Royal Holloway SVM http://svm.dcs.rhbnc.ac.uk
- 9. SVM World http://www.svmworld.com
- 10. Fraunhofer FIRST SVM page *http://svm.first.gmd.de*

#### **Presentation Bias**

Hypothesis: Order of presentation influences where users look, but not where they click!

"normal"	$ _{1}^{-}, _{2}^{-}$	$ _{1}^{+}, _{2}^{-}$	$ _{1}^{-}, _{2}^{+}$	$ _{1}^{+}, _{2}^{+}$	total
$rel(I_1) > rel(I_2)$	15	19	1	1	36
$rel(I_1) < rel(I_2)$	11	5	2	2	20
$rel(I_1) = rel(I_2)$	19	9	1	0	29
total	45	33	4	3	85
"swapped"	$ _{1}^{-}, _{2}^{-}$	$ _{1}^{+}, _{2}^{-}$	$ _{1}^{-}, _{2}^{+}$	$ _{1}^{+}, _{2}^{+}$	total
$rel(I_1) > rel(I_2)$	11	15	1	1	28
$rel(I_1) < rel(I_2)$	17	10	7	2	36
$rel(I_1) = rel(I_2)$	36	11	3	0	50
total	64	36	11	3	114

#### Quality-of-Context Bias

Hypothesis: Clicking depends only on the link itself, but not on other links.

	Rank of clicked link as sorted by relevance judges
Normal + Swapped	2.67
Reversed	3.27

=> Users click on less relevant links, if they are embedded between irrelevant links.

# Are Clicks Absolute Relevance Judgments?

- Clicks depend not only on relevance of a link, but also
  - On the position in which the link was presented
  - The quality of the other links
- => Interpreting Clicks as absolute feedback extremely difficult!

#### Strategies for Generating Relative Feedback

#### Strategies

- "Click > Skip Above" - (3>2), (5>2), (5>4)
- "Last Click > Skip Above" - (5>2), (5>4)
- "Click > Earlier Click" - (3>1), (5>1), (5>3)
- "Click > Skip Previous" - (3>2), (5>4)
- "Click > Skip Next" - (1>2), (3>4), (5>6)

- 1. Kernel Machines http://www.kernel-machines.org/
- 2. Support Vector Machine *http://jbolivar.freeservers.com/*
- 3. SVM-Light Support Vector Machine http://ais.gmd.de/~thorsten/svm light/
- 4. An Introduction to SVMs *http://www.support-vector.net/*
- 5. Support Vector Machine and ... http://svm.bell-labs.com/SVMrefs.html
- 6. Archives of SUPPORT-VECTOR... http://www.jisc.ac.uk/lists/SUPPORT...
- 7. Lucent Technologies: SVM demo applet *http://svm*.bell-labs.com/SVMsvt.html
- 8. Royal Holloway SVM *http://svm.dcs.rhbnc.ac.uk*
- 9. SVM World *http://www.svmworld.com*
- 10. Fraunhofer FIRST SVM page *http://svm.first.gmd.de*

#### **Comparison with Explicit Feedback**

Explicit Feedback	Abstracts
Data	Phase I
Strategy	"normal"
Inter-Judge Agreement	89.5
Click > Skip Above	$80.8\pm3.6$
Last Click > Skip Above	$83.1\pm3.8$
Click > Earlier Click	$67.2 \pm 12.3$
Click > Skip Previous	$82.3\pm7.3$
Click > No Click Next	$84.1\pm4.9$

=> All but "Click > Earlier Click" appear accurate

#### Is Relative Feedback Affected by Bias?

Explicit Feedback	Abstracts				
Data		Phase II			
Strategy	"normal"	"swapped"	"reversed"		
Click > Skip Above	$88.0 \pm 9.5$	79.6 ± 8.9	83.0 ± 6.7		
Last Click > Skip Above	$89.7 \pm 9.8$	$77.9 \pm 9.9$	$84.6 \pm 6.9$		
Click > Earlier Click	75.0 ± 25.8	36.8 ± 22.9	$28.6 \pm 27.5$		
Click > Skip Previous	$88.9 \pm 24.1$	$80.0 \pm 18.0$	$79.5 \pm 15.4$		
Click > No Click Next	75.6 ± 14.5	66.7 ± 13.1	70.0 ± 15.7		

⇒Significantly better than random in all conditions, except "Click > Earlier Click"

# How Well Do Users Judge Relevance Based on Abstract?

Explicit Feedback	Abstracts	Pages	
Data	Phase II		
Strategy	all	all	
Inter-Judge Agreement	82.5	86.4	
Click > Skip Above	$83.1 \pm 4.4$	$78.2 \pm 5.6$	
Last Click > Skip Above	$83.8 \pm 4.6$	$80.9 \pm 5.1$	
Click > Earlier Click	$46.9 \pm 13.9$	64.3 ±15.4	
Click > Skip Previous	$81.6\pm9.5$	$80.7 \pm 9.6$	
Click > No Click Next	$70.4\pm8.0$	$67.4 \pm 8.2$	

⇒clicks based on abstracts reflect relevance of the page well

# Learning Retrieval Functions from Pairwise Preferences

- Idea: Learn a ranking function, so that number of violated pair-wise training preferences is minimized.
- Form of Ranking Function: sort by  $U(q,d_i) = w_1 * (\#of query words in title of d_i) + w_2 * (\#of query words in anchor) + ... + w_n * (page-rank of d_i) = w * \Phi(q,d_i)$
- Training: Select w so that

if user prefers  $d_i$  to  $d_i$  for query q, then  $U(q, d_i) > U(q, d_i)$ 

#### **Ranking Support Vector Machine**

• Find ranking function with low error and large margin

$$nin \quad \frac{1}{2}\vec{w}\cdot\vec{w} + C\sum_{i,j,k}\xi_{kij}$$
  
s.t.  $\vec{w}\cdot\Phi(q_1,d_i) \ge \vec{w}\cdot\Phi(q_1,d_j) + 1 - \xi_{1ij}$   
...  
 $\vec{w}\cdot\Phi(q_n,d_i) \ge \vec{w}\cdot\Phi(q_n,d_j) + 1 - \xi_{nij}$ 

- Properties
  - Convex quadratic program
  - Non-linear functions using Kernels
  - Implemented as part of SVM-light
  - http://svmlight.joachims.org



#### Experiment

- Meta-Search Engine "Striver"
  - Implemented meta-search engine on top of Google, MSNSearch, Altavista, Hotbot, Excite
  - Retrieve top 100 results from each search engine
  - Re-rank results with learned ranking functions
- Experiment Setup
  - User study on group of ~20 German machine learning researchers and students
    - => homogeneous group of users
  - Asked users to use the system like any other search engine
  - Train ranking SVM on 3 weeks of clickthrough data
  - Test on 2 following weeks

# Which Ranking Function is Better? Balanced Interleaving

(u=tj, q="svm")

$f_1(u,q) \rightarrow r_1 \longleftarrow$		$\rightarrow$ f <sub>2</sub> (u,q) $\rightarrow$ r <sub>2</sub>
1.       Kernel Machines http://svm.first.gmd.de/         2.       Support Vector Machine http://jbolivar.freeservers.com/         3.       An Introduction to Support Vector Machines http://www.support-vector.net/         4.       Archives of SUPPORT-VECTOR-MACHINES http://www.jiscmail.ac.uk/lists/SUPPORT         5.       SVM-light Support Vector Machine http://ais.gmd.de/~thorsten/svm light/         Image: Support Vector Machine http://ois.gmd.de/~thorsten/svm light/         Image: Support Vector Machine http://ois.gmd.de/~thorsten/svm light/         Support Vector Machine http://ois.gmd.de/~thorsten/svm light/         Image: Support Vector Machine http://ois.gmd.de/~thorsten/svm light/         Support Vector Machine http://ois.gmd.de/~thorsten/svm light/         Image: Support Vector Machine http://www.jusport Vector Machine http://ois.gmd.de/~thorsten/svm light/         Image: Support Vector Machine http://support Vector Machine ht	Interleaving(r <sub>1</sub> ,r <sub>2</sub> )         . Kernel Machines         http://svm.first.gmd.de/         . Support Vector Machine         http://jbolivar.freeservers.com/         . SVM-Light Support Vector Machine         http://ais.gmd.de/~thorsten/svm light/         . An Introduction to Support Vector Machines         http://www.support Vector.net/         . Support Vector Machine and Kernel References         http://svm.research.bell-labs.com/SVM/refs.html         . Archives of SUPPORT-VECTOR-MACHINES         . Http://svm.research.bell-labs.com/SVT/SVMsvt.html	<ol> <li>Kernel Machines http://svm.first.gmd.de/</li> <li>SVM-Light Support Vector Machine http://ais.gmd.de/~thorsten/svm light/</li> <li>Support Vector Machine and Kernel References http://svm.research.bell-labs.com/SVMrefs.html</li> <li>Lucent Technologies: SVM demo applet http://svm.research.bell-labs.com/SVT/SVMsvt.html</li> <li>Royal Holloway Support Vector Machine http://svm.dcs.rhbnc.ac.uk</li> <li>Invariant: For all k, top k of balanced interleaving is union of top k<sub>1</sub> of r<sub>1</sub> and top k<sub>2</sub> of r<sub>2</sub> with k<sub>1</sub>=k<sub>2</sub> ± 1</li> </ol>

**Interpretation:**  $(r_1 \succ r_2) \leftrightarrow clicks(topk(r_1)) > clicks(topk(r_2))$ 

#### Results

Ranking A	Ranking B	A better	B better	Tie	Total
Learned	Google	29	13	27	69
Learned	MSNSearch	18	4	7	29
Learned	Toprank	21	9	11	41

#### Result:

- Learned > Google
- Learned > MSNSearch
- Learned > Toprank

Toprank: rank by increasing minimum rank over all 5 search engines

### Learned Weights

•	Weight	Feature
•	0.60	cosine between query and abstract
•	0.48	ranked in top 10 from Google
•	0.24	cosine between query and the words in the URL
•	0.24	doc ranked at rank 1 by exactly one of the 5 engines
•		
•	0.22	host has the name "citeseer"
•		
•	0.17	country code of URL is ".de"
•	0.16	ranked top 1 by HotBot
•	•••	
•	-0.15	country code of URL is ".fi"
•	-0.17	length of URL in characters
•	-0.32	not ranked in top 10 by any of the 5 search engines
•	-0.38	not ranked top 1 by any of the 5 search engines

#### Conclusions

- Clickthrough data can provide accurate feedback
  - Clickthrough provides relative instead of absolute judgments
- Ranking SVM can learn effectively from relative preferences
  - Improved retrieval through personalization in meta search
- Current and future work
  - Exploiting query chains
  - Other implicit feedback signals
  - Adapting intranet search for ArXiv.org
  - Recommendation
  - Robustness to "click-spam"
  - Learning and micro-economic theory for interactive learning with preference
  - Further user studies to get better models of user behavior

#### Feedback across Query Chains

🖉 MSN Search: svm - Microsoft Internet Explorer		
File Edit View Favorites Tools Help	MSN Search: support vector machine - Microsoft Internet Explorer	<u>- 🗆 ×</u>
😋 Back 🔻 🕥 👻 📓 🐔 🔎 Search 👷 Favorites  🖉 🗢 🌺 🔯 🔻 🖵 🎎 🐴	File Edit View Favorites Tools Help	
Address 🚳 http://search.msn.com/results.aspx?q=svm&FORM=QBHP	😮 Back 🔻 🕤 👻 😰 🏠 🔎 Search 🔹 Favorites  🔗 🕏 😓 💽 👻 🗒 🖏	
Google - msn search - G Search - 🛷 PageBank 🕸 37 blocked	Address 💩 http://search.msn.com/results.aspx?q=support+vector+machine&FORM=Ql 🔻 🎅 Go	Links »
<u>Web News Images Desktop Encarta</u>	Coogle ▼ msn search ▼ C Search ▼ 🚿 PageBank 🖾 37 blocked 🌾 Check ▼ »	
svm 🗨 Near Me	Web News Images Desktop Encarta	
+Search Builder Serving: Help Español	support vector machine Search 🔻 Near Me	
Web Results reformulate	+Search Builder Settings Help Español	
1-10 of 220,590 containing <b>svm</b> (0.14 seconds)	Web Results	
Buy SVM Stock for \$4 - www.sharebuilder.com	1-10 of 63,199 containing support vector machine (0.22 seconds)	
No account or investment minimums and no inactivity fee. Automatically build a diversified portfo	Programming Vester File Format Sunnert, unulgediade com	
ServiceMaster: In-depth Company Info - www.hoovers.com	C/C++, VB, Delphi, Net programmers: Create vector imaging software with support for loading, editing, proces	sing, s
Go to Hoover's Online for in-depth, inst-hand, company coverage provided by business expens.	Support Vector Machines - analytics.infotrack.net	
ServiceMaster We Are Home	Learn all about genetic programming in terms and contexts you can understand.	
ServiceMaster Reports First Quarter 2005 Results ServiceMaster Announces Secon	Buy "Support Vector Machines" at BN.com - www.barnesandnoble.com	
www.svm.com Cached page 6/12/2005	Buy "Support vector Machines" by Lipo wang at Barnes & Noble. Fast and tree delivery. Three days or less on	orders
SVM-World de	Support Vector Machines - The Book - Support Vector	SP
am Freitag Nachmittag auf den letzten Drücker gegen weiterlesen Willkommen	AN INTRODUCTION TO <b>SUPPORT VECTOR</b> MACHINES (and other kernel-based learning	<u>Supp</u>
größten deutsch- sprachigen Fan-Homepages vom SV Meppen. Hier wirst du stets m	78019 5 NEWS: School	Machi Shon
www.svm-word.de <u>Cacheo page</u> 6/12/2005	www.support-vector.net Cached page	everyti
SVM sri od orocione costiera La conosità di gostire la immogini finica dei sistemi predetti d	Support Vector Machine - The Software	specia www.e
monitoraggio dei fenomeni di erosione e dei lavori di ripascimento costiero	on recent advances in statistical learning theory. This page gives pointers to free	Sunn
www.svm.it <u>Cached page</u>	software about the authors	Amaz
School of Volunteer Management	www.support-vector.net/software.html Cached page	Buy bo
The School of Volunteer Management offers a range of volunteer management and	Show more results from "www.support-vector.net".	shopp
www.svm.net.au Cached page	Support vector machine - Wikipedia, the free encyclopedia	WWW.8
SV Mattersburg Online	the distance to the nearest cleanly split examples. This work popularized the expression Support Vector Machine or SVM . The SVM was popularized in the machine learning	<u>See y</u>
www.svm.at <u>Cached page</u>	community by Bernhard Schölkopf	
	en.wikipedia.org/wiki/Support_Vector_Machine Cached page	
🖉	GIST: Support Vector Machine 1.0 - Data submission	