On Filtering the Web

Web Information and Quality Evaluation 2010

Nedim Lipka - nedim.lipka@uni-weimar.de

Valencia, September 14, 2010

Outline

On Filtering the Web

Web Information and Quality Evaluation 2010

Nedim Lipka - nedim.lipka@uni-weimar.de

Valencia, September 14, 2010

Outline

On Filtering the Web

Web Information and Quality Evaluation 2010

Nedim Lipka - nedim.lipka@uni-weimar.de

Valencia, September 14, 2010

Outline

Information Filtering

Information Filtering

Characteristics

Information Retrieval (IR)

- A user has an information need.
- A query (imperfectly) represents the information need.
- An IR-system is typically used in a one-time fashion.

Information Filtering (IF)

- Groups or individuals have regular information interests.
- A profile or a query represents a regular information interests.
- An IF-system is typically used repeatedly by persons with long-term goals.

Information Filtering

Tasks

Filtering issues: . . . spam, quality, fraud, authorship, genres, topics, sentiment, humor, language, gender, writing-styles. . .

Information Filtering

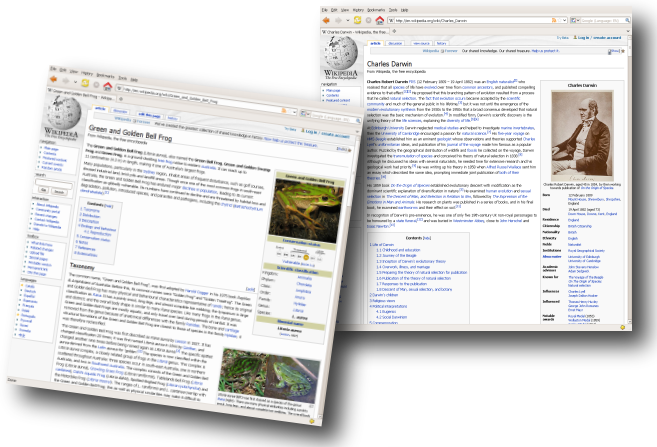
Example: Identifying featured articles in Wikipedia

Filtering is sometimes challenging. E.g. identifying featured articles in Wikipedia, where a featured article is:

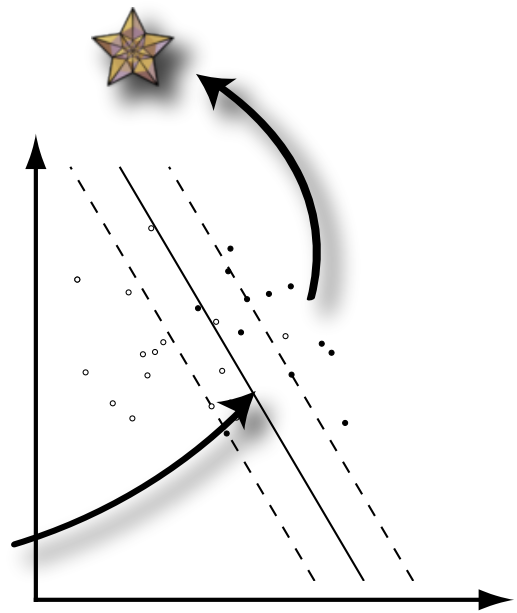
Information Filtering

Filtering with Machine Learning

Overview.



$[0,0.01,\ldots,0.05]^T$



Information Filtering

Representations

Our Web-based plagiarism analysis application takes the suspicious docu...

Information Filtering

Word 3-grams

Our Web-based plagiarism analysis application takes the suspicious docu...

Information Filtering

Word 3-grams

Information Filtering

Word 3-grams

Information Filtering

Word 3-grams

Our Web-based plagiarism analysis application takes the suspicious docu...

Information Filtering

Part of Speech 2-grams

Information Filtering

Part of Speech 2-grams

Information Filtering

Part of Speech 2-grams

Information Filtering

Part of Speech 2-grams

Information Filtering

A Content- and Style-based Representation: Character 5-grams

Our Web-based plagiarism analysis application takes the suspicious docu...

Information Filtering

A Content- and Style-based Representation: Character 5-grams

O u r W e b - b a s e d p l a g i a r i s m a n a l y s i s a p p l i c a t i o n t a k e s t h e s u s p i c i o u s d o c u . . .

Information Filtering

A Content- and Style-based Representation: Character 5-grams

Our Web-based plagiarism analysis application takes the suspicious docu...

Information Filtering

A Content- and Style-based Representation: Character 5-grams

Our Web-based plagiarism analysis application takes the suspicious docu...

Information Filtering

Example: Identifying featured articles in Wikipedia

Setting.

- Extracted Plaintexts from English Wikipedia.
- 180 featured / 180 non-featured biology articles.
- 200 featured / 200 non-featured history articles.

Information Filtering

Example: Identifying featured articles in Wikipedia

Results.

Representation	**Classifier**	**Identification of featured articles (P/R/F)**	
Cross Validation.		*within Biology*	*within History*
bin char trigram	SVM	0.966 / 0.961 / **0.964**	0.888 / 0.955 / **0.920**
bin POS trigram	SVM	0.949 / 0.933 / 0.941	0.889 / 0.925 / 0.907
word count	SVM	0.755 / 0.600 / 0.669	0.874 / 0.870 / 0.872
bag of words	NB	0.832 / 0.989 / 0.904	0.860 / 0.950 / 0.903
Domain Transfer.		*History* $\rightarrow$ *Biology*	*Biology* $\rightarrow$ *History*
bin char trigram	SVM	0.800 / 0.978 / **0.880**	0.886 / 0.855 / **0.870**
bin POS trigram	SVM	0.799 / 0.883 / 0.839	0.898 / 0.790 / 0.840
word count	SVM	0.772 / 0.733 / 0.752	0.878 / 0.830 / 0.853
bin bag of words	SVM	0.800 / 0.889 / 0.842	0.930 / 0.665 / 0.776

Information Filtering

Example: Identifying featured articles in Wikipedia

The most discriminative character trigrams.

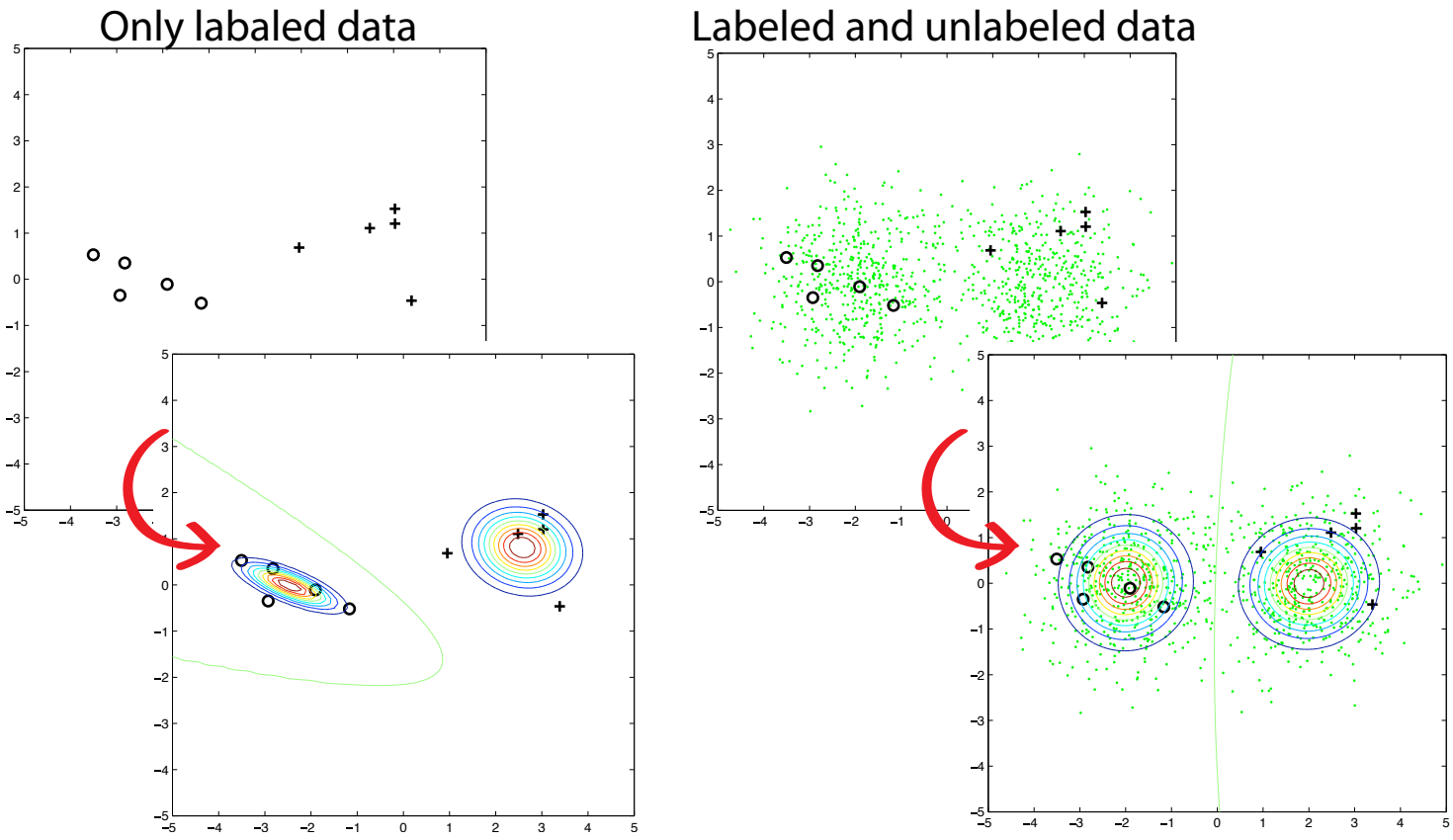
ing	ng_	,_a	at_	e,_	er_	_an	ed_	d_a
_be	ter	s_a	_re	as_	ted	g_a	tha	n_t
_a_	ly_	to_	_th	nd_	._a	on_	sed	t_t

transitions affixes
punctuation stopwords

Exploiting Unlabeled Data

Exploiting Unlabeled Data

Can unlabeled data be useful?



Exploiting Unlabeled Data

The Semi-Supervised Smoothness Assumption

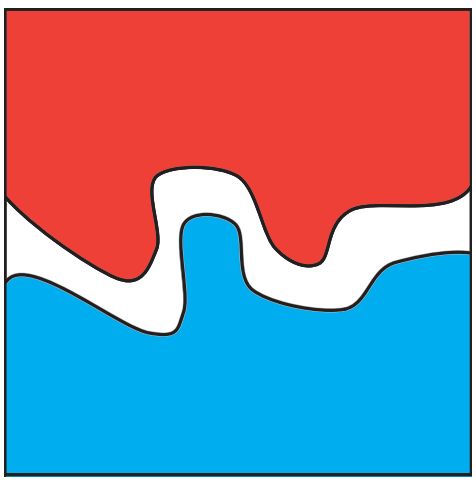
"If two points x_1, x_2 in a high-density region are close, then so should be the corresponding outputs y_1, y_2."

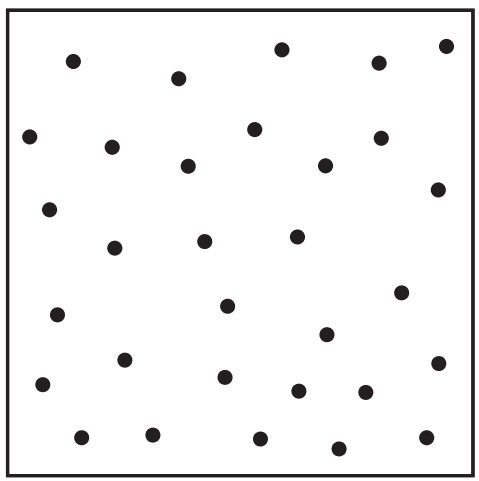
Exploiting Unlabeled Data

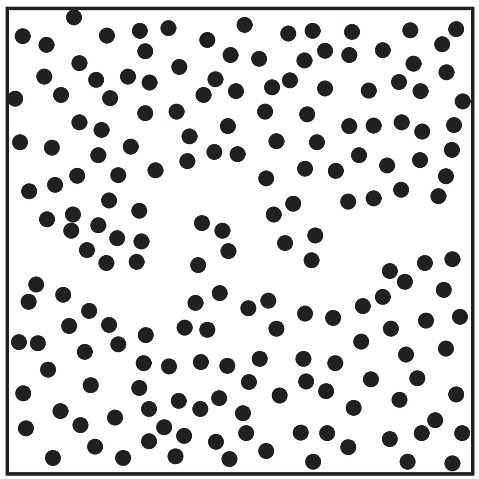
The Cluster Assumption

"If points are in the same cluster, they are likely to be of the same class."

"The decision boundary should lie in a low-density region."





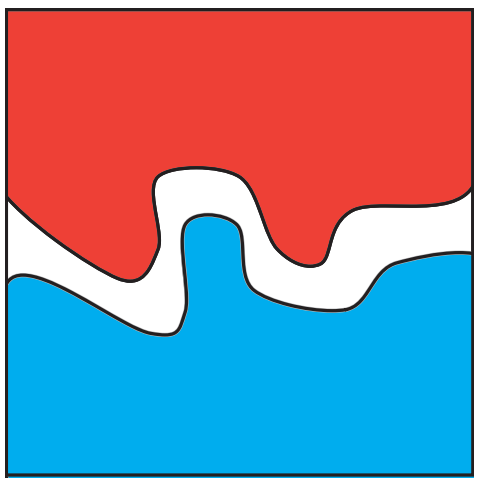


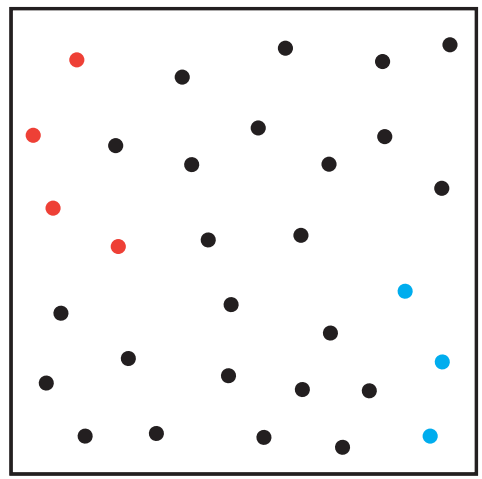
Exploiting Unlabeled Data

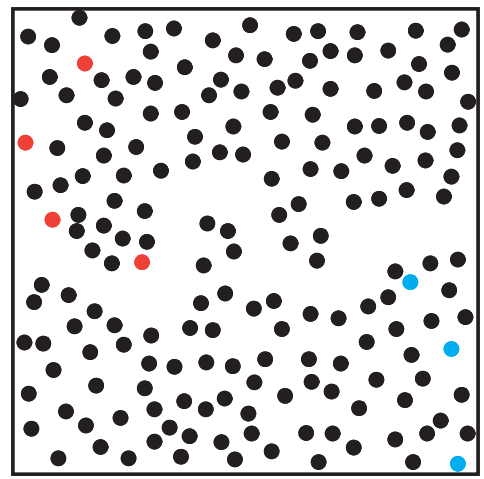
The Cluster Assumption

"If points are in the same cluster, they are likely to be of the same class."

"The decision boundary should lie in a low-density region."





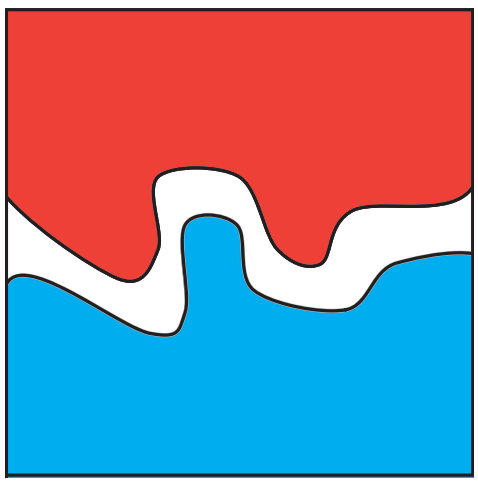


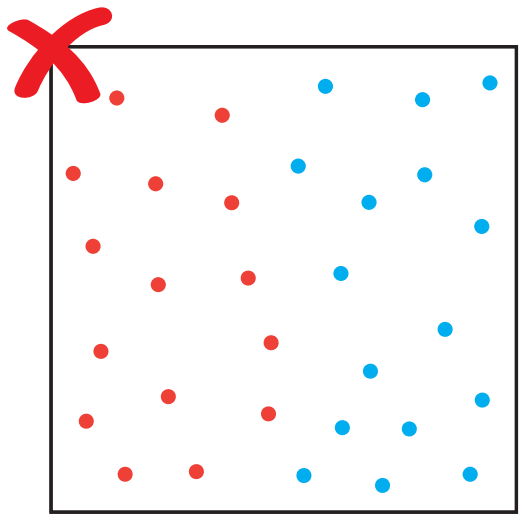
Exploiting Unlabeled Data

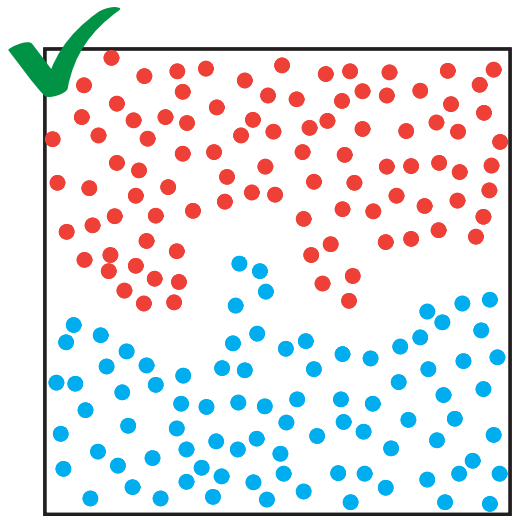
The Cluster Assumption

"If points are in the same cluster, they are likely to be of the same class."

"The decision boundary should lie in a low-density region."



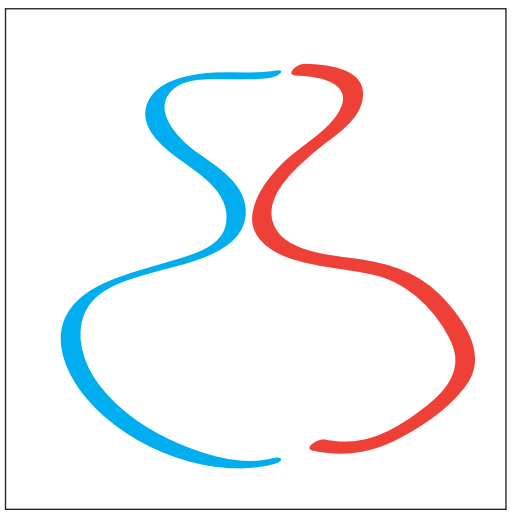


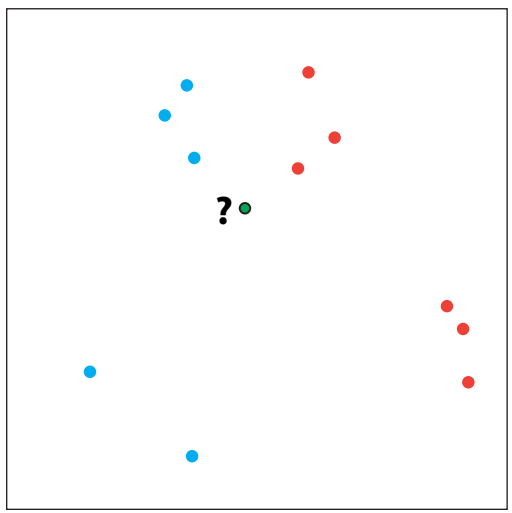


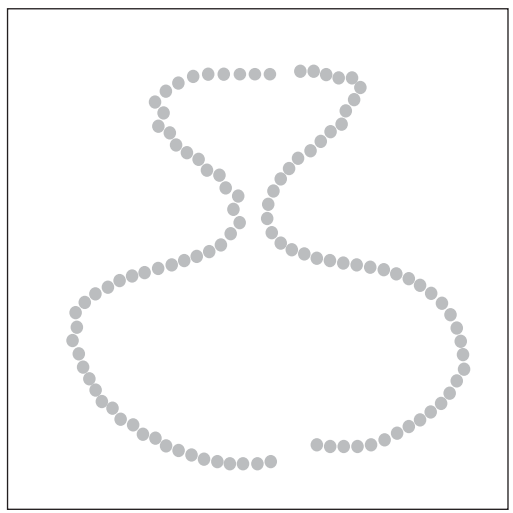
Exploiting Unlabeled Data

The Manifold Assumption

"The (high-dimensional) data lie (roughly) on a low-dimensional manifold."



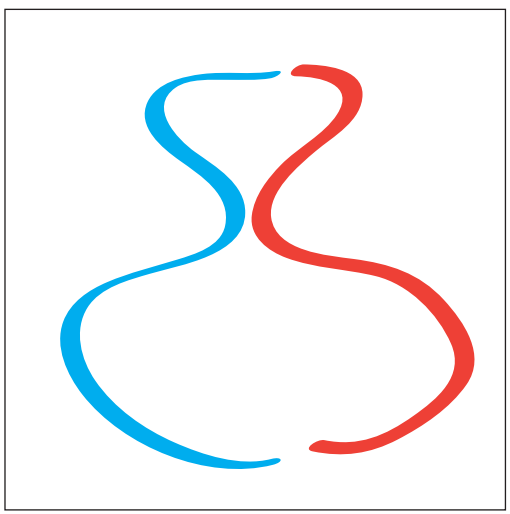


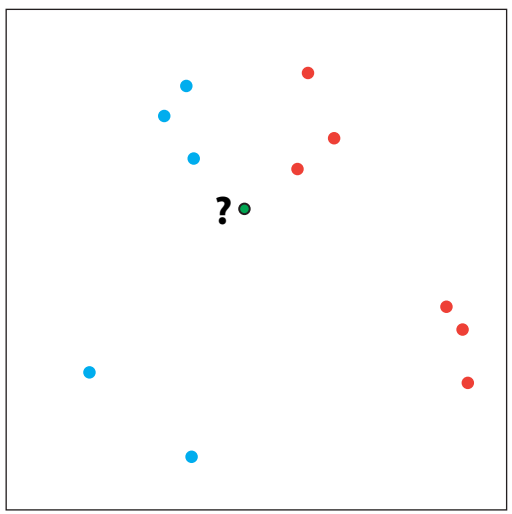


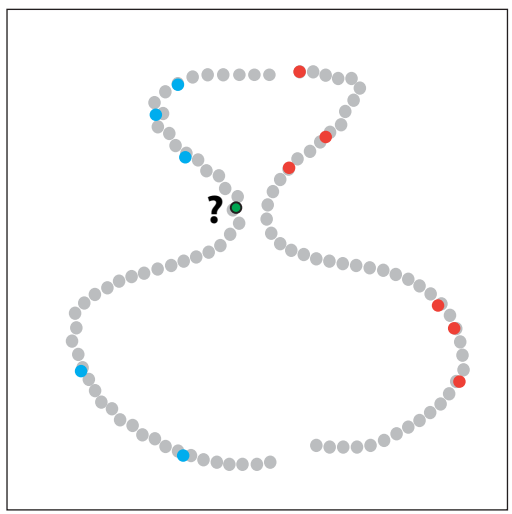
Exploiting Unlabeled Data

The Manifold Assumption

"The (high-dimensional) data lie (roughly) on a low-dimensional manifold."







Exploiting Unlabeled Data

Co-Training

Goal:

- Extend the labeled training set.

Requirements:

- Unlabeled data.
- Two representations (views).

Exploiting Unlabeled Data

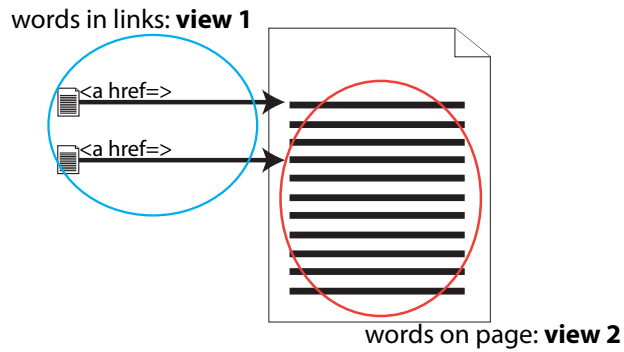
Co-Training

Goal:

- Extend the labeled training set.

Requirements:

- Unlabeled data.
- Two representations (views).



Exploiting Unlabeled Data

Co-Training: Algorithm

Input:

- Labeled training examples L
- Unlabeled examples U

Create a sampling pool $U' \subset U$. For k iterations:

1. Train classifier h_1, h_2 with L considering representation $\mathbf{x_1}, \mathbf{x_2}$ of $\mathbf{x}$.
2. Classify U' with h_1, remove p positive and n negative examples with the highest confidence and add them to L.
3. Classify U' with h_2, remove p positive and n negative examples with the highest confidence and add them to L.
4. Randomly choose $2p + 2n$ examples from U to replenish U'.

Exploiting Unlabeled Data

Co-Training: Constraints

1. Each view should be sufficient for correct classification.

Exploiting Unlabeled Data

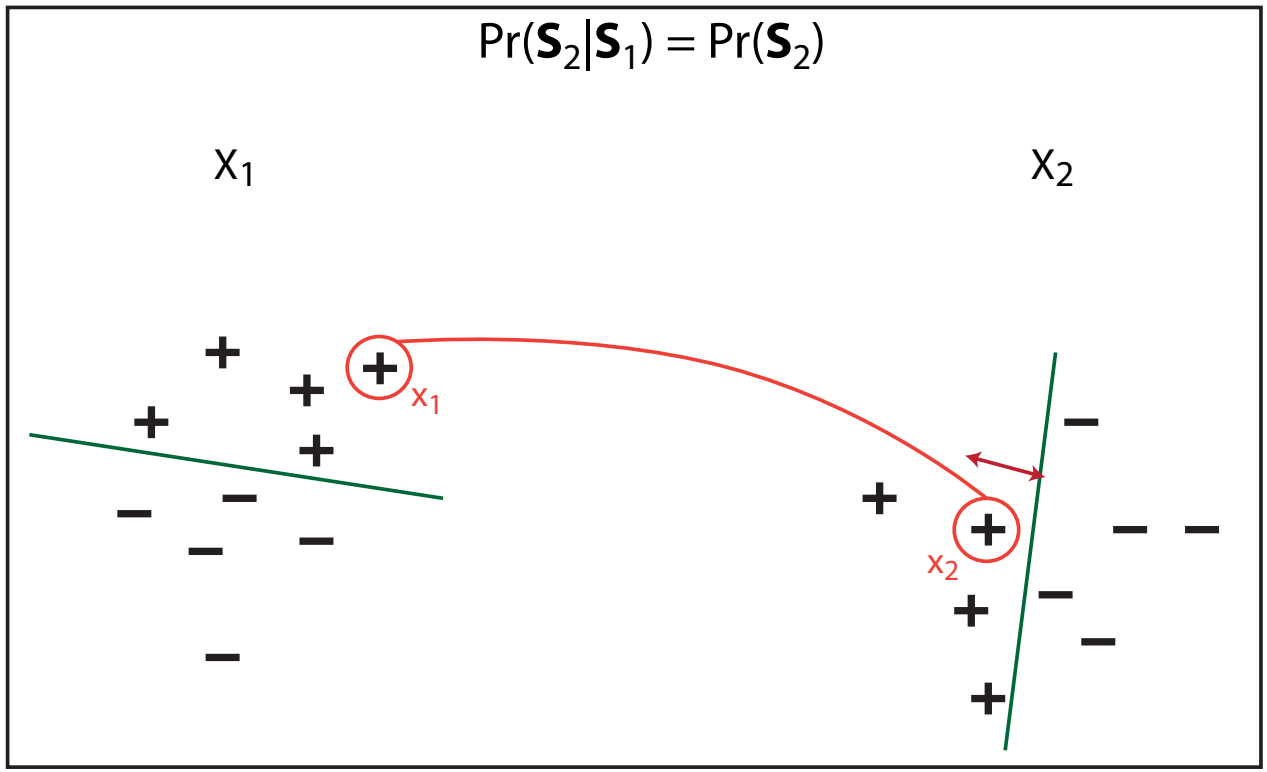
Co-Training: Constraints

1. Each view should be sufficient for correct classification.
2. Independence assumptions on the representations...

Exploiting Unlabeled Data

Co-Training: Constraints

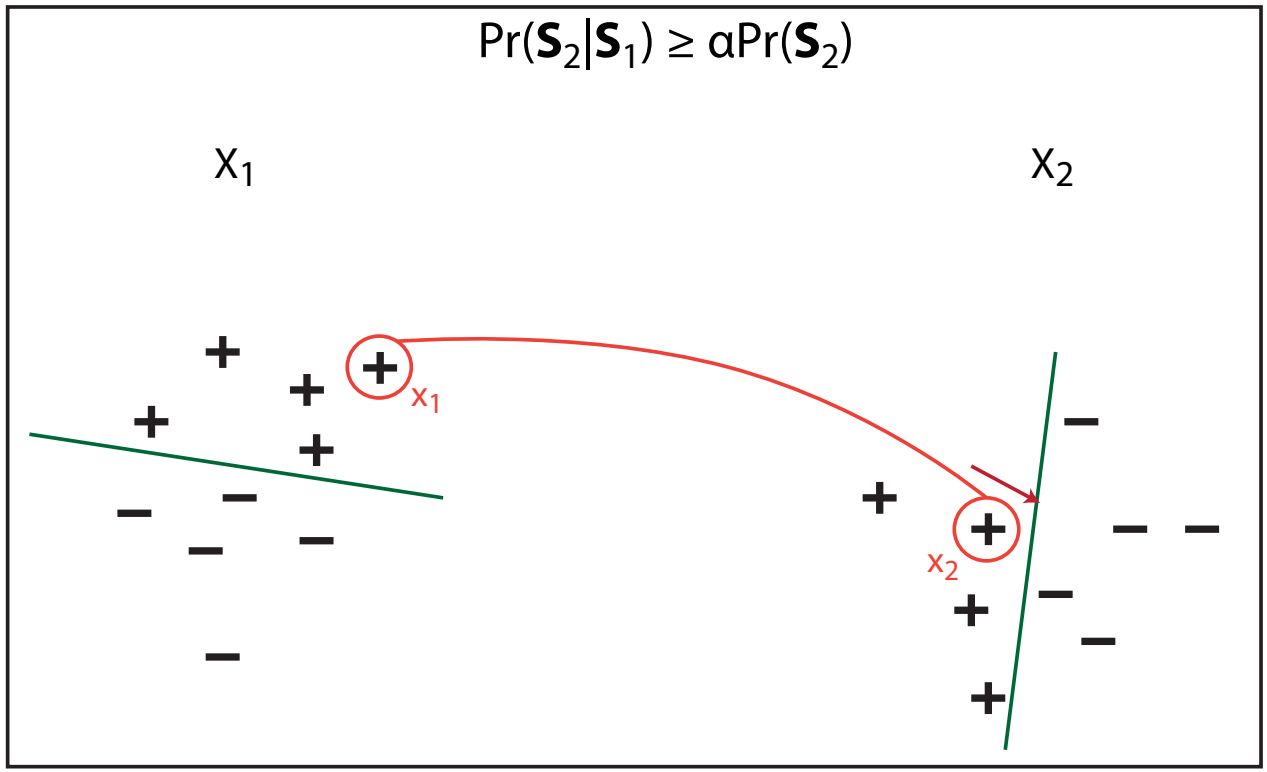
- class-conditional independence



Exploiting Unlabeled Data

Co-Training: Constraints

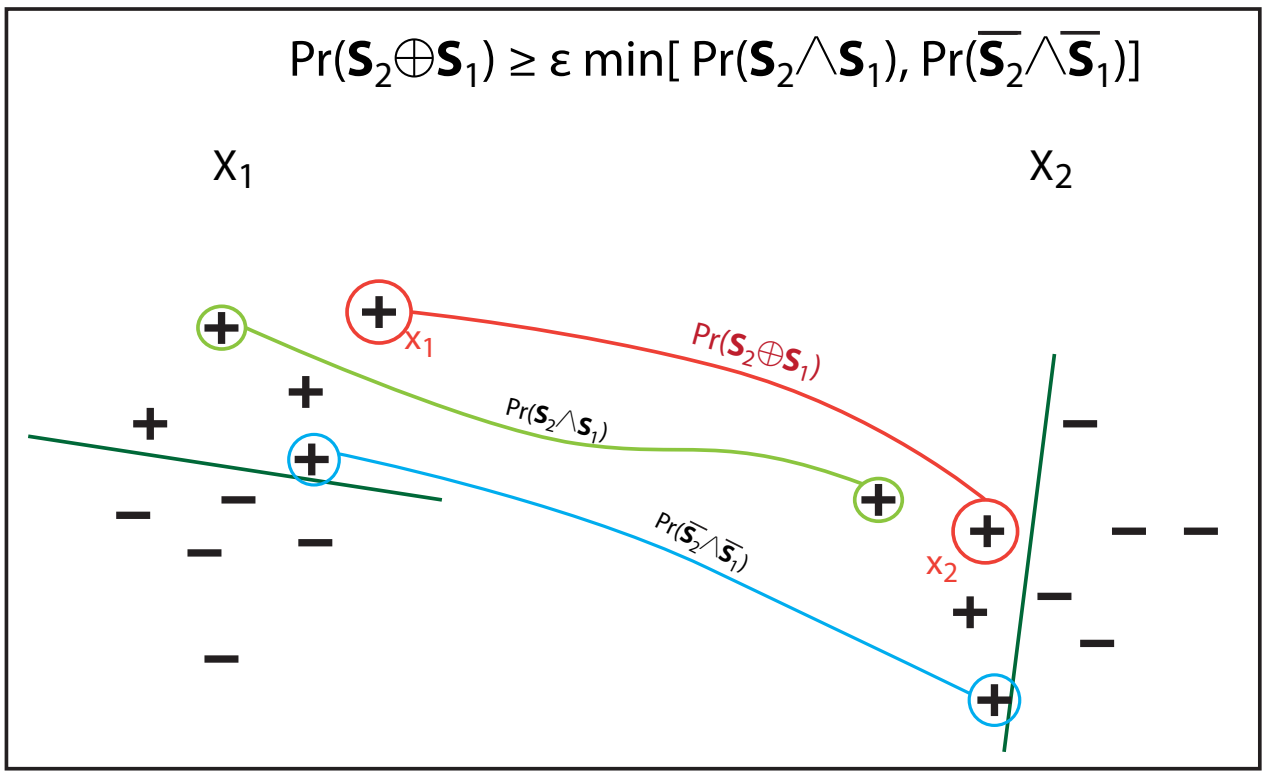
- class-conditional independence
- weak dependence



Exploiting Unlabeled Data

Co-Training: Constraints

- class-conditional independence
- weak dependence
- ϵ-expansion



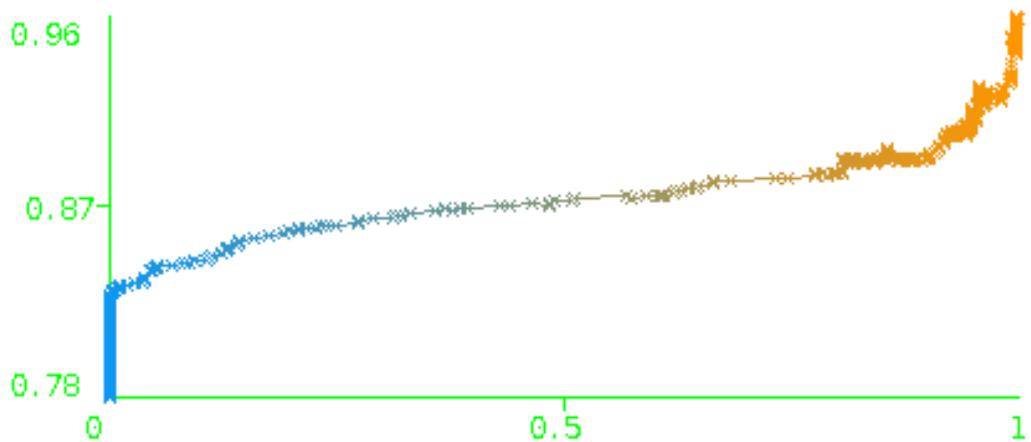
Tradeoff: classifier performance - potential for improvement

Exploiting Unlabeled Data

Co-Training: Can a the learning algorithm benefit from self-labeled data?

"... remove p positive and n negative **examples with the highest confidence** and add them to L."

The precision at a high confidence ought to be high in order to compile a valuable training set.



Precision - Confidence

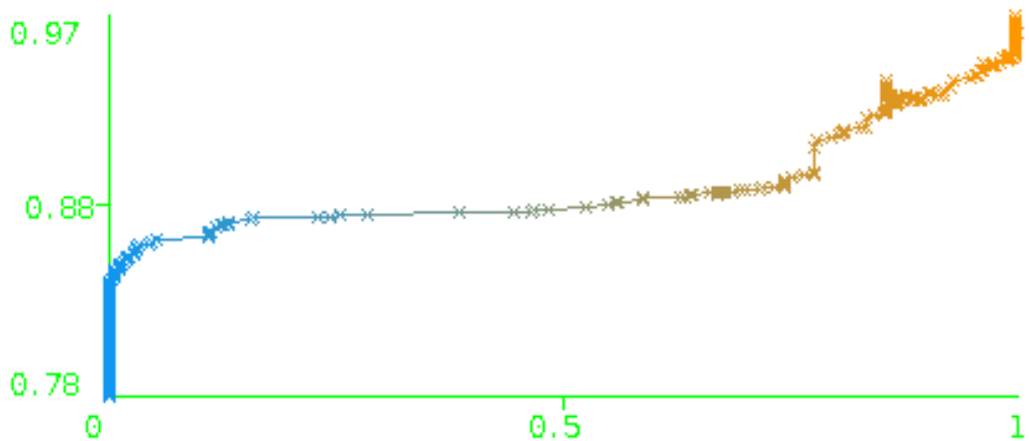
90% training data (900 documents): 86, 3 acc. (inlinks) - 92, 4 acc. (page)

Exploiting Unlabeled Data

Co-Training: Can a the learning algorithm benefit from self-labeled data?

"... remove p positive and n negative **examples with the highest confidence** and add them to L."

The precision at a high confidence ought to be high in order to compile a valuable training set.



Precision - Confidence

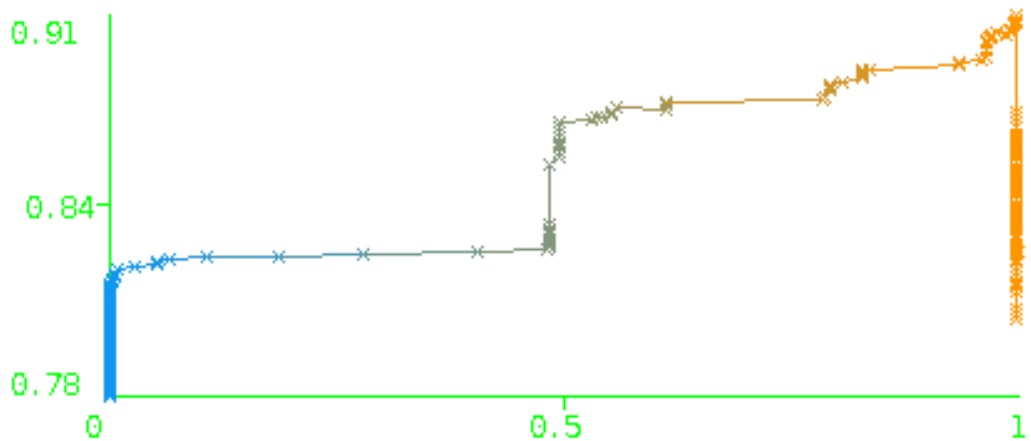
50% training data (500 documents): 86, 5 acc. (inlinks) - 90, 7 acc. (page)

Exploiting Unlabeled Data

Co-Training: Can a the learning algorithm benefit from self-labeled data?

"... remove p positive and n negative **examples with the highest confidence** and add them to L."

The precision at a high confidence ought to be high in order to compile a valuable training set.



Precision - Confidence

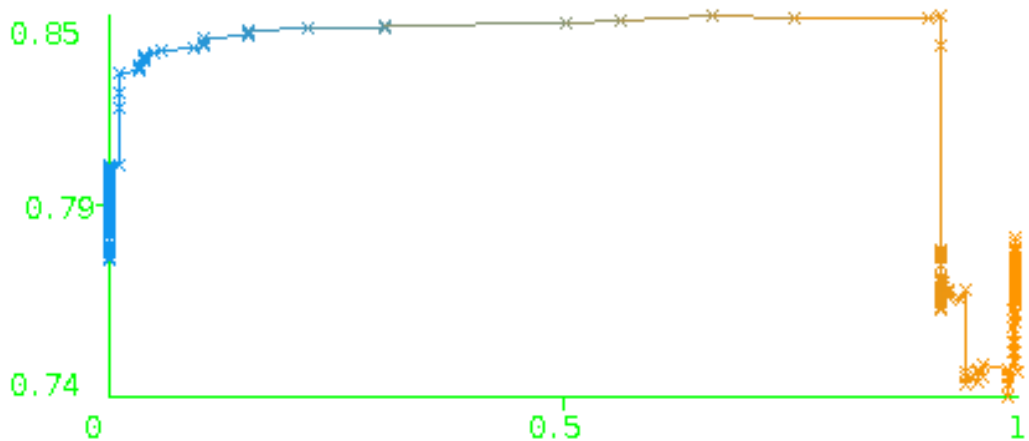
10% training data (100 documents): 87, 4 acc. (inlinks) - 91, 4 acc. (page)

Exploiting Unlabeled Data

Co-Training: Can a the learning algorithm benefit from self-labeled data?

"... remove p positive and n negative **examples with the highest confidence** and add them to L."

The precision at a high confidence ought to be high in order to compile a valuable training set.



Precision - Confidence

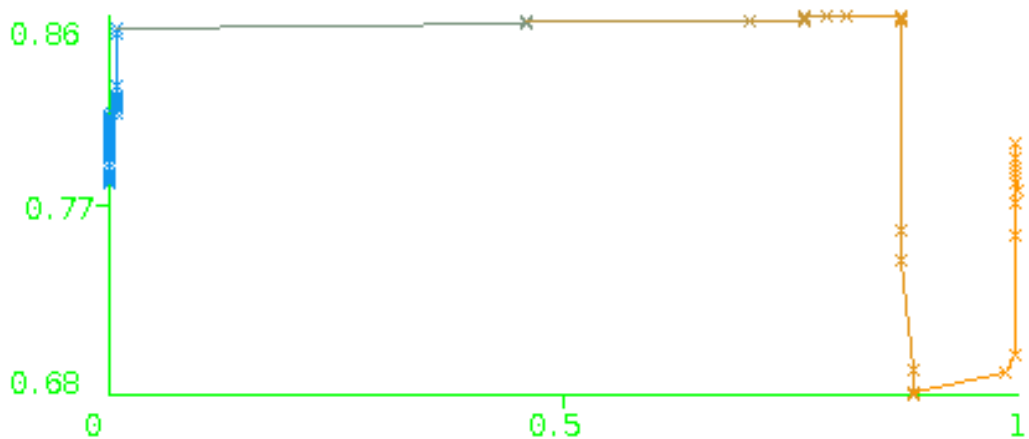
5% training data (50 documents): 84, 6 acc. (inlinks) - 90, 7 acc. (page)

Exploiting Unlabeled Data

Co-Training: Can a the learning algorithm benefit from self-labeled data?

"... remove p positive and n negative **examples with the highest confidence** and add them to L."

The precision at a high confidence ought to be high in order to compile a valuable training set.



Precision - Confidence

1% training data (10 documents): 86, 3 acc. (inlinks) - 69, 6 acc. (page)

Exploiting Unlabeled Data

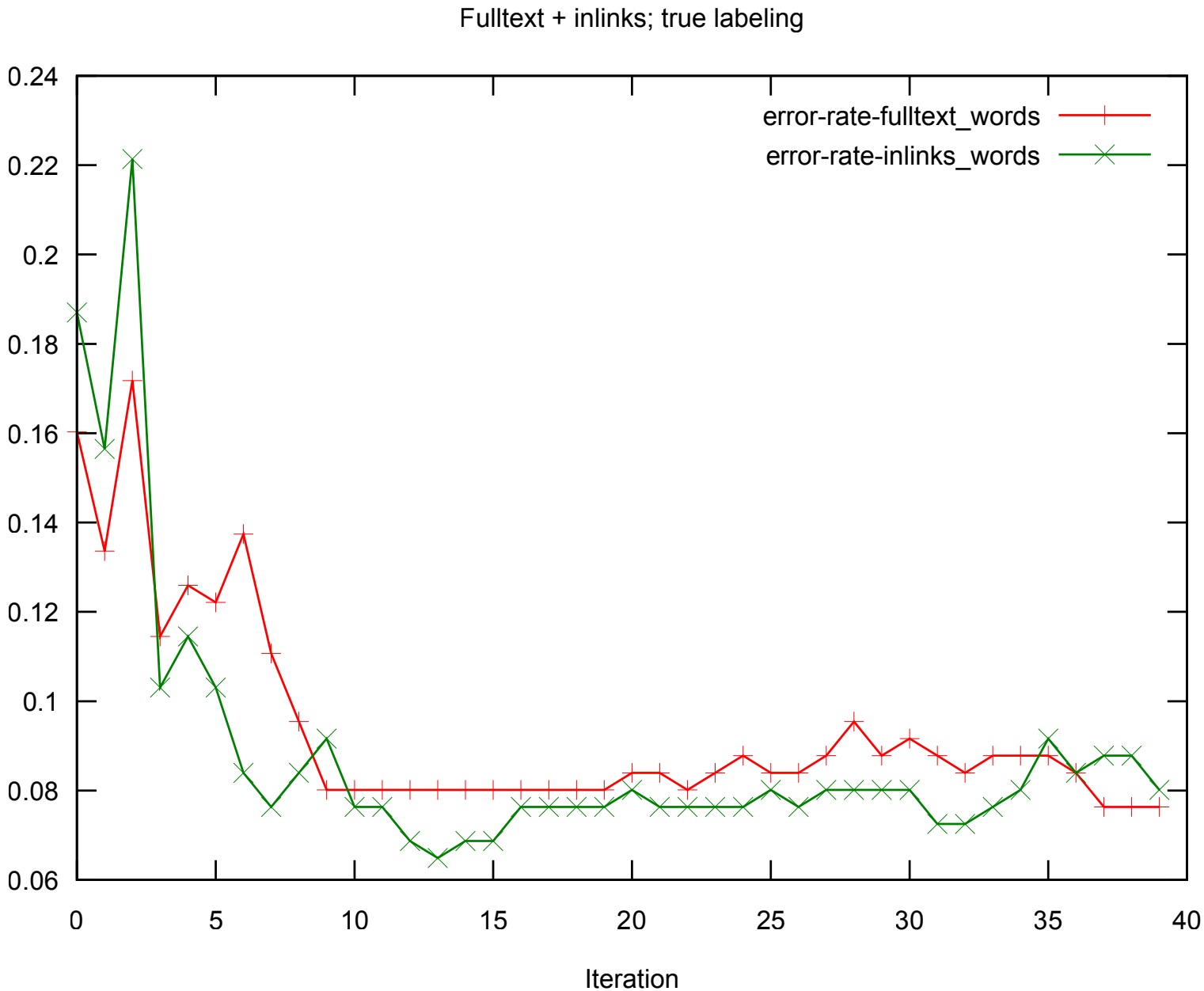
Co-Training: Experiment

Setting.

- Corpus with 230 course and 821 non-course webpages.
- Selection of 3 positive and 9 negative examples in each iteration.
- Initial sampling pool U' contains 75 examples.

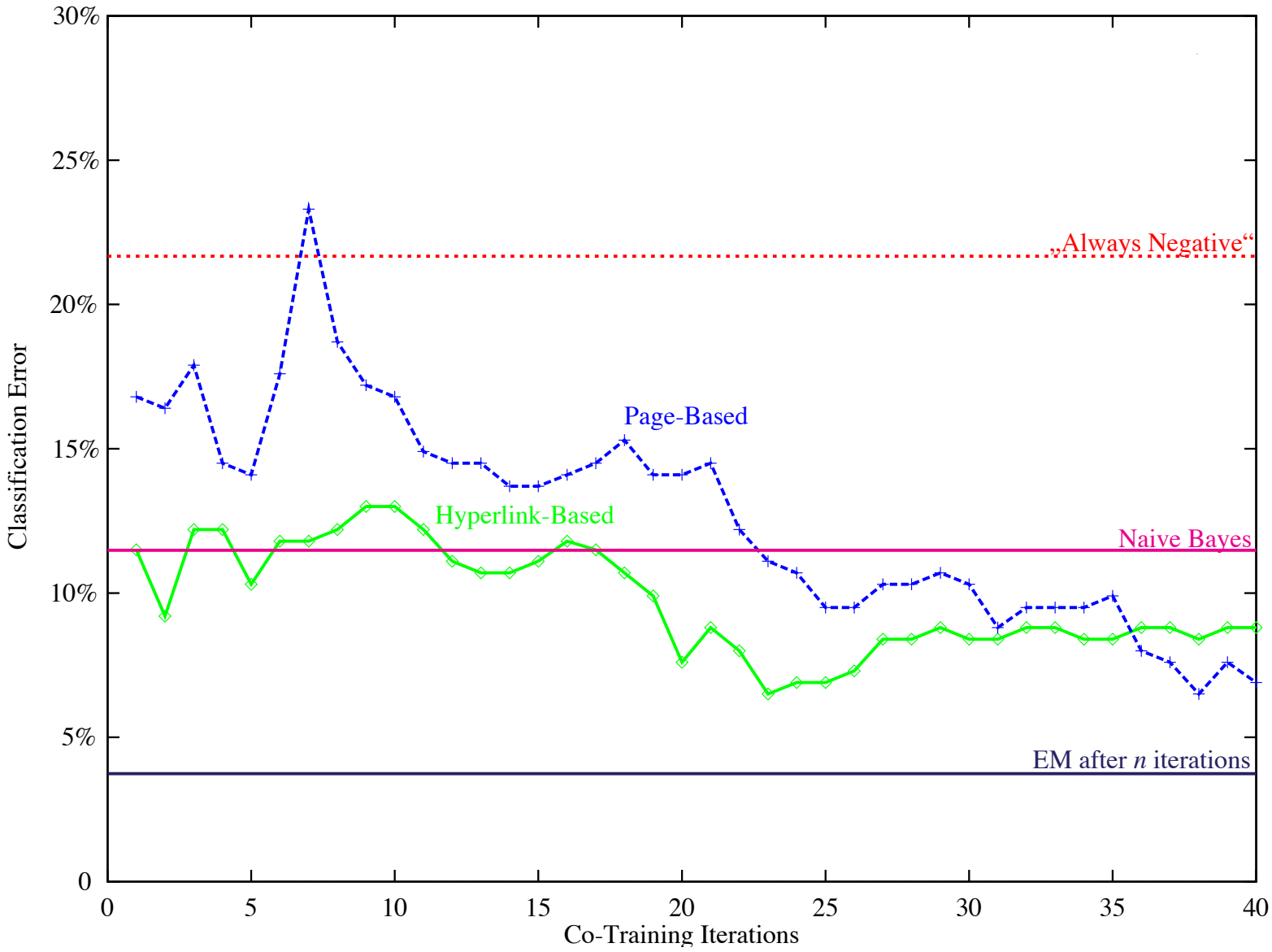
Exploiting Unlabeled Data

Co-Training: Experiment



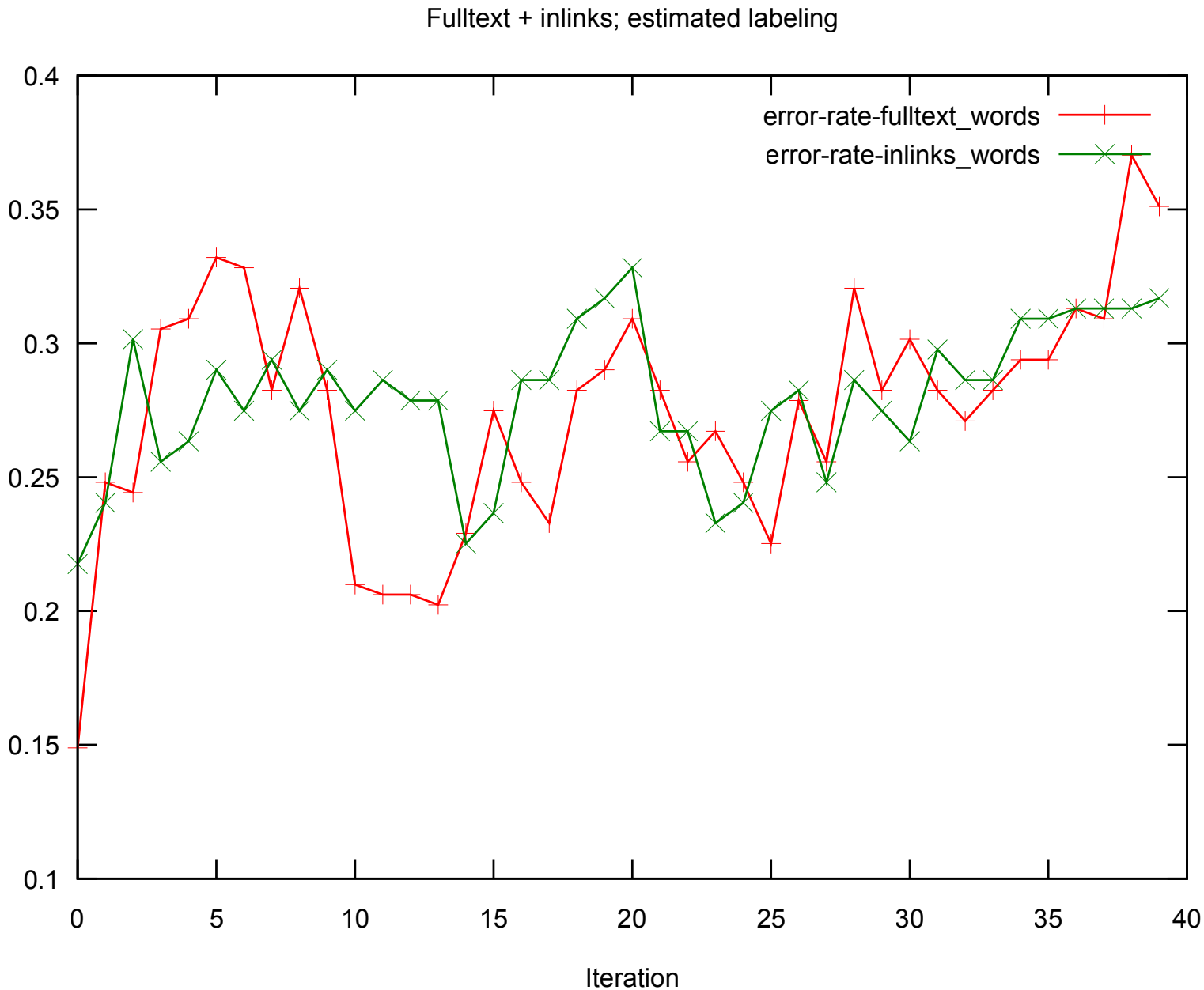
Exploiting Unlabeled Data

Co-Training: Experiment



Exploiting Unlabeled Data

Co-Training: Experiment



La Sinopsis

Remember:
When trying to solve some problem, one should not solve a more difficult problem as an intermediate step.