Exploratory analysis of time-varying word embeddings

Syrielle MONTARIOL

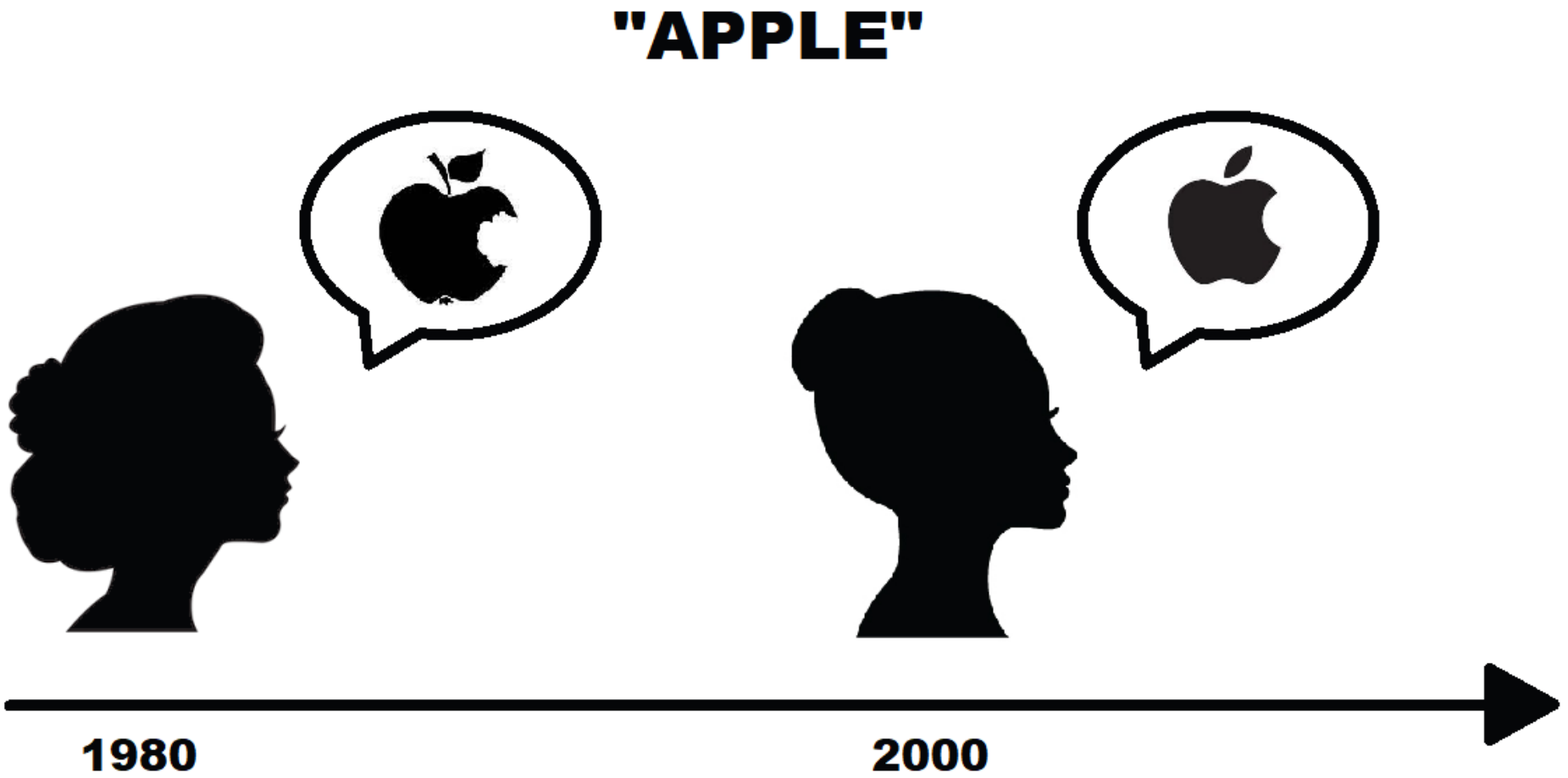
Supervisor: **Alexandre ALLAUZEN**

LIMSI / TLP - ED STIC

CIFRE Ph.D with Société Générale

Journée Des Doctorants du LIMSI - 06/06/2019

An example of evolution



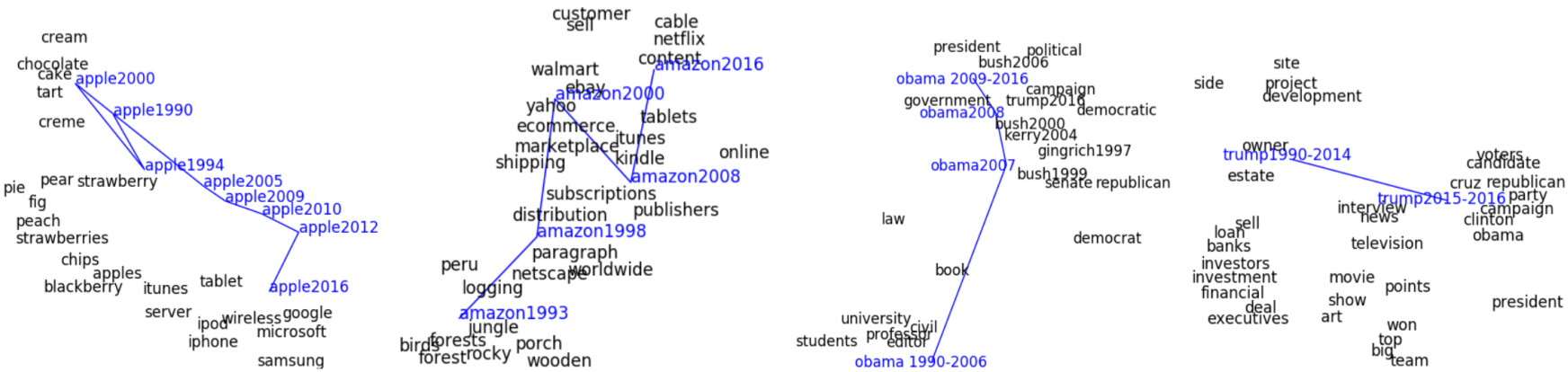
An example of short-term cross-lingual evolution

How to detect this?

Detecting evolution of word embedding

Hypothesis : a change in the context of a word reflects a change of its meaning.

- Word use, meaning or connotation change ; new word appear and disappear ; new sense appear and disappear.
- Tool: Time-varying word embeddings



(a) apple (b) amazon (c) obama (d) trump

Figure 1: Trajectories of brand names and people through time: `apple`, `amazon`, `obama`, and `trump`.

Literature I

Divide the corpus into T time slices : $t_1, t_2, ..., t_T$.

Incremental updating [Kim et al., 2014]

- Learn an embedding for the first time slice
- Initialize the algorithm of successive time slices with the previous one (*incremental update*).

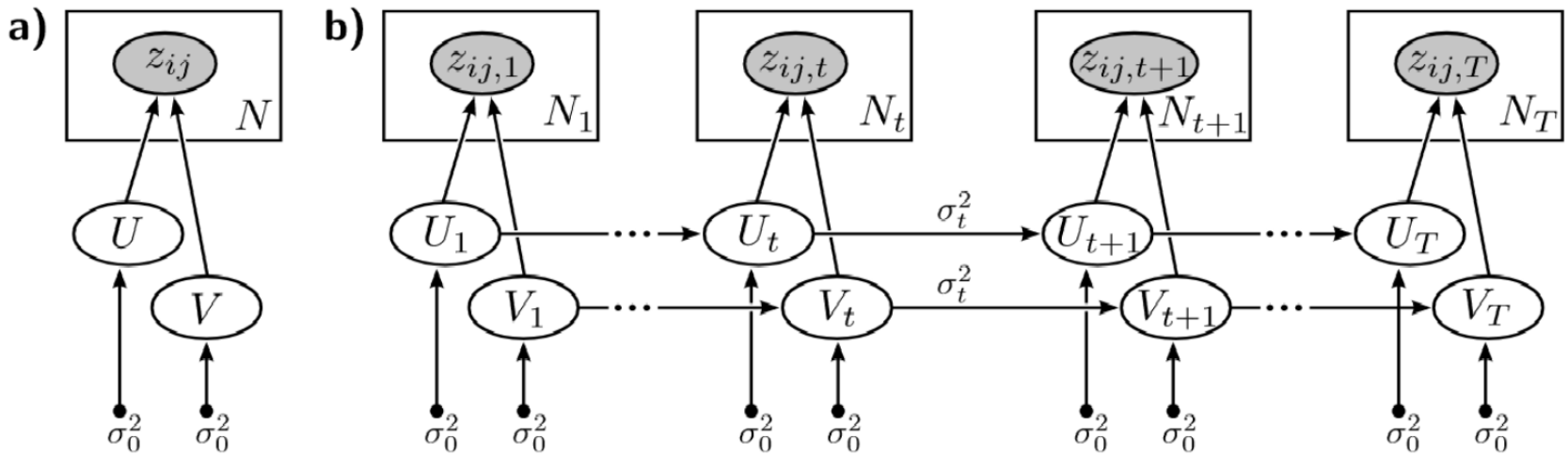
Alignment of embedding spaces [Hamilton et al., 2016, Kulkarni et al., 2015]

- Learn one embedding for each time slice;
- Alignment of the vectorial spaces (with optimization methods : find the best linear transformation to align across time periods)
- Assumption: the meaning of most words is stable $\Rightarrow$ when the alignment model fails for one word, it may indicate a drift.

Literature II

Probabilistic methods [Bamler and Mandt, 2017, Rudolph and Blei, 2018]

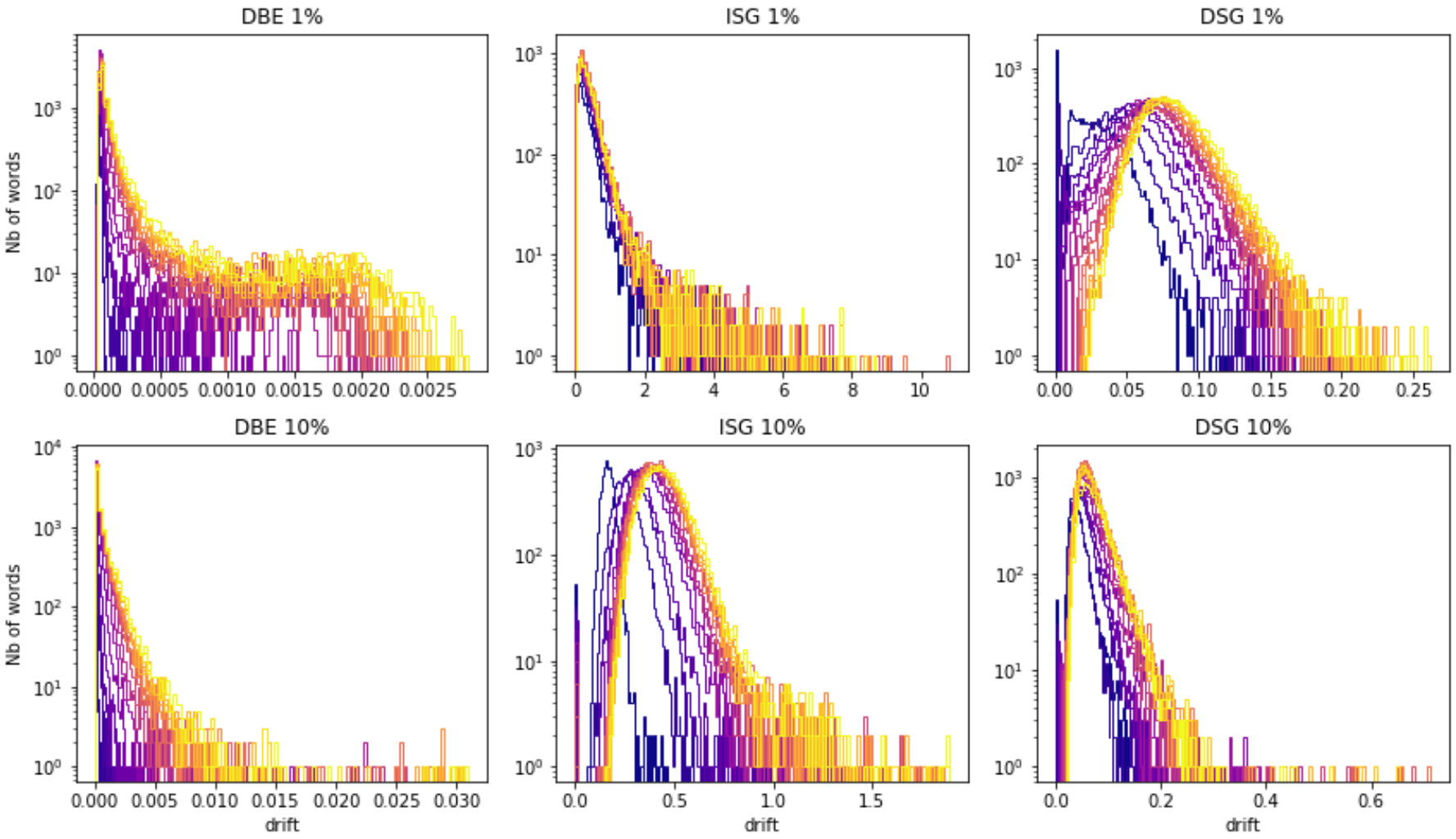
- Jointly learn the probabilistic embeddings and a latent diffusion process across all time slices
- The diffusion process controls the drift of the embeddings



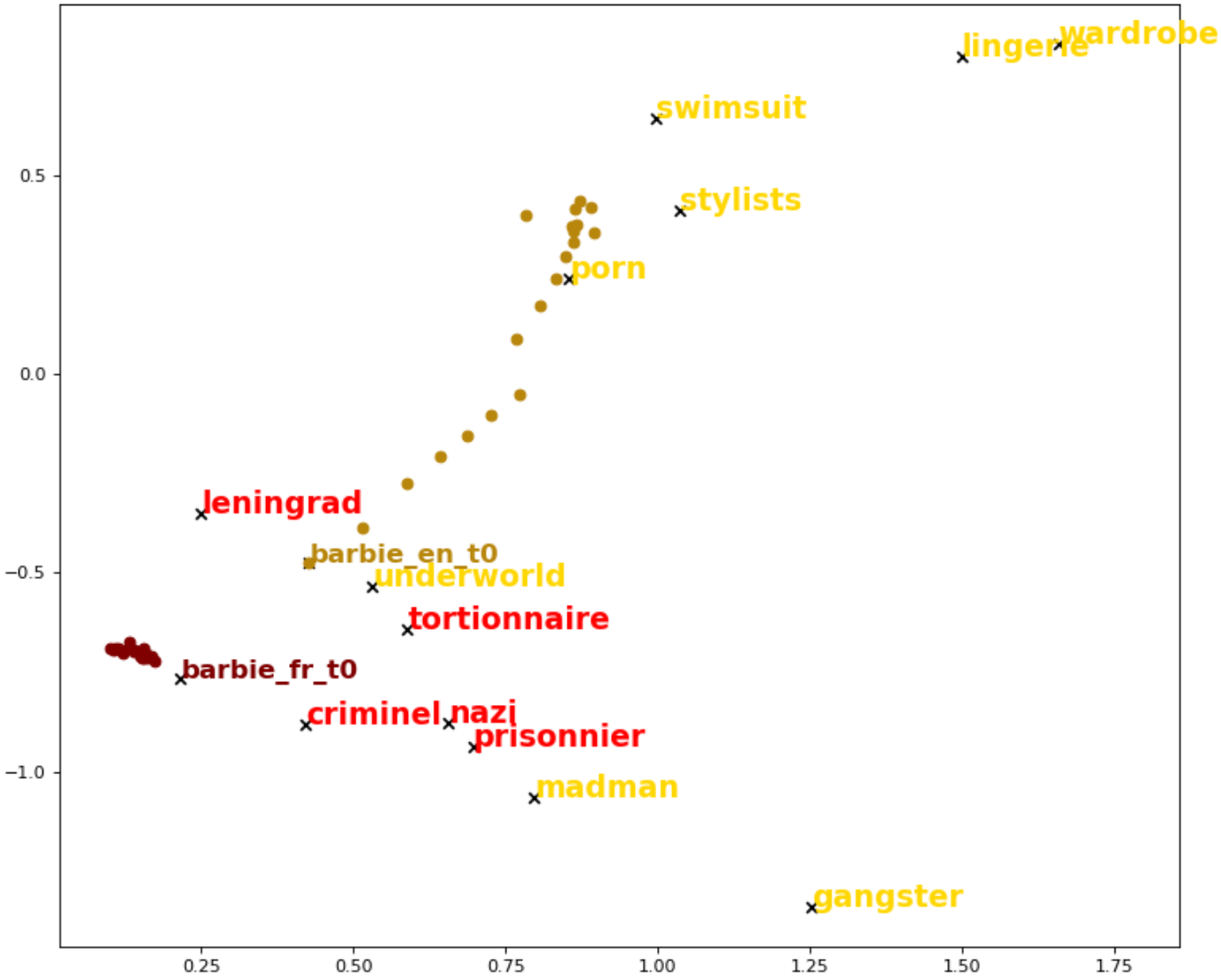
Advantage against static methods

- Smoother trajectories, Need less data per time step
- Easier to disambiguate random noise from semantic shift

Drift distribution for scarce data



Example of drift in two languages



Challenges of the field

Evaluation and comparison

- Vast amount of historical data, no large manually labeled corpus
- No ground truth
- Lack of standardized evaluation practices
- No standard experimental process $\Rightarrow$ few comparison between methods.

Language is complex

- Word polysemy : Simplification in most works = vector representation of a word $\sim$ its *main* sense.
- Need sense-aware algorithms to quantify the evolution of the weights of the sens of a word.
- No individual and independent shift ! Words shift together in a correlated way. See it as a network of word relations.

Applications

How can we use it ?

- Short-term : event detection, topics popularity...
- Long-term : Evolution of languages (= Diachrony), global law of semantic shift...
- Enhancing accuracy for tasks on a temporal corpus (Information Retrieval, ...)

Future work

- Diachronic **sentiment-specific** embeddings
- For short-term reputation risk
- Using label propagation

References I

Bamler, R. and Mandt, S. (2017).
Dynamic word embeddings.
In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 380–389, International Convention Centre, Sydney, Australia. PMLR.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016).
Diachronic word embeddings reveal statistical laws of semantic change.
In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501. Association for Computational Linguistics.

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., and Petrov, S. (2014).
Temporal analysis of language through neural language models.
In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65. Association for Computational Linguistics.

References II

Kulkarni, V., Al-Rfou, R., Perozzi, B., and Skiena, S. (2015).
Statistically significant detection of linguistic change.
In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 625–635, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Rudolph, M. and Blei, D. (2018).
Dynamic embeddings for language evolution.
In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1003–1011. International World Wide Web Conferences Steering Committee.