Structural and Visual Similarity Learning for Web Page Archiving

Marc Teva Law Carlos Sureda Gutierrez Nicolas Thome Stéphane Gançarski
Matthieu Cord

LIP6, UPMC - Sorbonne University, Paris, France
FirstName.LastName@lip6.fr

Abstract

We present in this paper a Web page archiving approach combining image and structural techniques. Our main goal is to learn a similarity between Web pages in order to detect whether successive versions of pages are similar or not. Our system is based on a visual similarity measure designed for Web pages. Combined with a structural analysis of Web page source codes, a supervised feature selection method adapted to Web archiving is proposed. Experiments on real Web archives are reported including scalability issues.

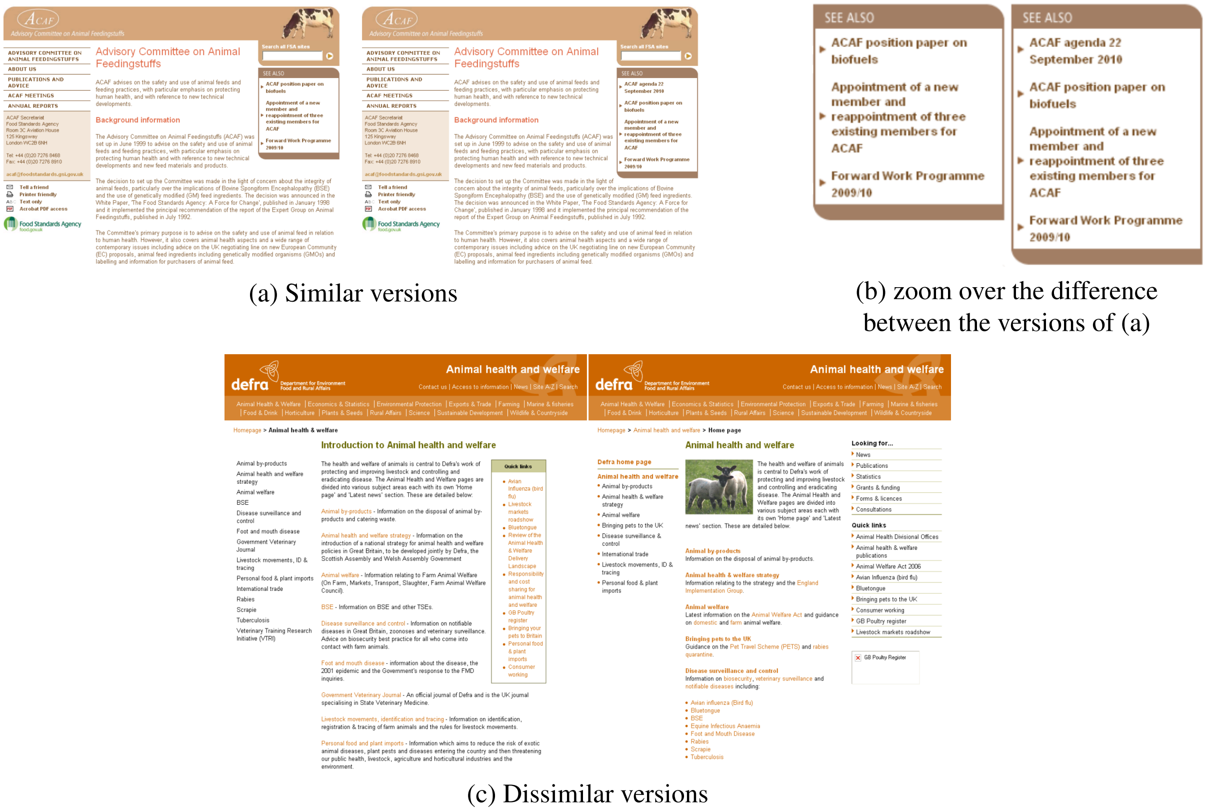
1 Introduction

The Web has become the most important way to spread information. Archiving the Web is thus crucial to preserve some useful information for future generations of historians, researchers, writers, ..., or citizens. Most of the time, Web archiving is performed by Web crawlers (bots) that capture Web pages and the associated media (e.g. images, videos...). To update archives, crawlers have to regularly revisit Web pages, but they generally do not know if or when changes appeared.

Archivists are facing a great challenge to maintain the quality of collected data (that should reflect the real Web): optimizing the crawling so that new versions are captured and/or kept only when changes are important while limiting the loss of useful information. A way to optimize crawling is to estimate the behaviour of a site in order to guess when or with which frequency it must be visited, and thus to study the importance of changes between successive versions [4]. For instance, the insertion of an advertisement link, illustrated in Fig 1(a) and Fig 1(b), is not related to the main information shared by the Web page. In contrast, changes in Fig 1(c) are significant. The crawling of the second version was thus necessary. In this paper, similarity functions for Web page comparison are investigated.

Most archivists only take into account the Web page source code (code string, DOM tree...) [13] and not the visual rendering [4,5,14]. However, the code may not be sufficient to describe the content of Web pages, e.g. images are usually defined only by their URL addresses, or scripts may be coded in different languages that make them hard to compare. Ben Saad et al. [4] propose to use the tree obtained by running the VIPS [5] algorithm on the rendered Web page. They obtain a rich semantic segmentation into blocks and then estimate a function of the importance of changes between page versions by comparing the different blocks. The VIPS structure of a Web page is a segmentation tree based on its DOM tree. It detects visual structures in the rendering of a page (e.g. tables) and tries to keep nodes (blocks) as homogeneous as possible. Two successive paragraphs without html tags will tend to be kept in the same node, whereas table elements with different background colors will be separated in different nodes. Image processing methods have been proposed for Web page segmentation. Cao et al. [6] preprocess the rendering of Web pages by an edge detection algorithm, and iteratively divide zones until all blocks are indivisible. They do not take the source code of pages into account. In the context of phishing detection, Fu et al. [9] compute similarities between Web pages using color and spatial visual feature vectors. However, they are only interested in the detection of exact copies.

We investigate in this paper structural and visual features and similarities to carry out an efficient page comparison system for Web archiving. To get a similarity close to archivists' attempts, both structural and visual informations are useful: structural to catch the dissimilarity if different scripts have the same rendering or if the hyperlinks are changed, visual if the codes of the versions of a Web page are unchanged but a loaded image was updated. Additionally, we propose a machine learning framework to set all the similarity parameters and combination weights. Our contribution is three-fold: (1) a complete hybrid Web page comparison framework combining computer vision and structural comparison methods, (2) a new measure dedicated to Web archiving that only considers the visible part of pages



ACAF

Advisory Committee on Animal Feedingstuffs

ADVISORY COMMITTEE ON ANIMAL FEEDINGSTUFFS

ABOUT US

PUBLICATIONS AND ADVICE

ACAF MEETINGS

ANNUAL REPORTS

ACAF Secretariat
Food Standards Agency
Room 3C Aviation House
125 Kingsway
London WC2B 6NH

Tel: +44 (0)20 7276 8468
Fax: +44 (0)20 7276 8910

acaf@foodstandards.gsi.gov.uk

✉ Tell a friend
Printer friendly
ABC Text only
PDF Acrobat PDF access

Food Standards Agency
food.gov.uk

Advisory Committee on Animal Feedingstuffs

ACAF advises on the safety and use of animal feeds and feeding practices, with particular emphasis on protecting human health, and with reference to new technical developments.

Background information

The Advisory Committee on Animal Feedingstuffs (ACAF) was set up in June 1999 to advise on the safety and use of animal feeds and feeding practices, with particular emphasis on protecting human health and with reference to new technical developments and new feed materials and products.

The decision to set up the Committee was made in the light of concern about the integrity of animal feeds, particularly over the implications of Bovine Spongiform Encephalopathy (BSE) and the use of genetically modified (GM) feed ingredients. The decision was announced in the White Paper, 'The Food Standards Agency: A Force for Change', published in January 1998 and it implemented the principal recommendation of the report of the Expert Group on Animal Feedingstuffs, published in July 1992.

The Committee's primary purpose is to advise on the safety and use of animal feed in relation to human health. However, it also covers animal health aspects and a wide range of contemporary issues including advice on the UK negotiating line on new European Community (EC) proposals, animal feed ingredients including genetically modified organisms (GMOs) and labelling and information for purchasers of animal feed.

Search all FSA sites

SEE ALSO

- ACAF position paper on biofuels
- Appointment of a new member and reappointment of three existing members for ACAF
- Forward Work Programme 2009/10

ACAF

Advisory Committee on Animal Feedingstuffs

ADVISORY COMMITTEE ON ANIMAL FEEDINGSTUFFS

ABOUT US

PUBLICATIONS AND ADVICE

ACAF MEETINGS

ANNUAL REPORTS

ACAF Secretariat
Food Standards Agency
Room 3C Aviation House
125 Kingsway
London WC2B 6NH

Tel: +44 (0)20 7276 8468
Fax: +44 (0)20 7276 8910

acaf@foodstandards.gsi.gov.uk

✉ Tell a friend
Printer friendly
ABC Text only
PDF Acrobat PDF access

Food Standards Agency
food.gov.uk

Advisory Committee on Animal Feedingstuffs

ACAF advises on the safety and use of animal feeds and feeding practices, with particular emphasis on protecting human health, and with reference to new technical developments.

Background information

The Advisory Committee on Animal Feedingstuffs (ACAF) was set up in June 1999 to advise on the safety and use of animal feeds and feeding practices, with particular emphasis on protecting human health and with reference to new technical developments and new feed materials and products.

The decision to set up the Committee was made in the light of concern about the integrity of animal feeds, particularly over the implications of Bovine Spongiform Encephalopathy (BSE) and the use of genetically modified (GM) feed ingredients. The decision was announced in the White Paper, 'The Food Standards Agency: A Force for Change', published in January 1998 and it implemented the principal recommendation of the report of the Expert Group on Animal Feedingstuffs, published in July 1992.

The Committee's primary purpose is to advise on the safety and use of animal feed in relation to human health. However, it also covers animal health aspects and a wide range of contemporary issues including advice on the UK negotiating line on new European Community (EC) proposals, animal feed ingredients including genetically modified organisms (GMOs) and labelling and information for purchasers of animal feed.

Search all FSA sites

SEE ALSO

- ACAF agenda 22 September 2010
- ACAF position paper on biofuels
- Appointment of a new member and reappointment of three existing members for ACAF
- Forward Work Programme 2009/10

(a) Similar versions

SEE ALSO

- ACAF position paper on biofuels
- Appointment of a new member and reappointment of three existing members for ACAF
- Forward Work Programme 2009/10

SEE ALSO

- ACAF agenda 22 September 2010
- ACAF position paper on biofuels
- Appointment of a new member and reappointment of three existing members for ACAF
- Forward Work Programme 2009/10

(b) zoom over the difference between the versions of (a)

defra Department for Environment Food and Rural Affairs

Animal health and welfare

Contact us | Access to information | News | Site A-Z | Search

Animal Health & Welfare | Economics & Statistics | Environmental Protection | Exports & Trade | Farming | Marine & fisheries | Food & Drink | Horticulture | Plants & Seeds | Rural Affairs | Science | Sustainable Development | Wildlife & Countryside

Homepage > **Animal health & welfare**

Animal by-products
Animal health and welfare strategy
Animal welfare
BSE
Disease surveillance and control
Foot and mouth disease
Government Veterinary Journal
Livestock movements, ID & tracing
Personal food & plant imports
International trade
Rabies
Scrapie
Tuberculosis
Veterinary Training Research Initiative (VTRI)

Introduction to Animal health and welfare

The health and welfare of animals is central to Defra's work of protecting and improving livestock and controlling and eradicating disease. The Animal Health and Welfare pages are divided into various subject areas each with its own 'Home page' and 'Latest news' section. These are detailed below:

Animal by-products - Information on the disposal of animal by-products and catering waste.

Animal health and welfare strategy - Information on the introduction of a national strategy for animal health and welfare policies in Great Britain, to be developed jointly by Defra, the Scottish Assembly and Welsh Assembly Government

Animal welfare - Information relating to Farm Animal Welfare (On Farm, Markets, Transport, Slaughter, Farm Animal Welfare Council).

BSE - Information on BSE and other TSEs.

Disease surveillance and control - Information on notifiable diseases in Great Britain, zoonoses and veterinary surveillance. Advice on biosecurity best practice for all who come into contact with farm animals.

Foot and mouth disease - information about the disease, the 2001 epidemic and the Government's response to the FMD inquiries.

Government Veterinary Journal - An official journal of Defra and is the UK journal specialising in State Veterinary Medicine.

Livestock movements, identification and tracing - Information on identification, registration & tracing of farm animals and the rules for livestock movements.

Personal food and plant imports - Information which aims to reduce the risk of exotic animal diseases, plant pests and diseases entering the country and then threatening our public health, livestock, agriculture and horticultural industries and the environment.

Quick links

- Avian Influenza (bird flu)
- Livestock markets roadshow
- Bluetongue
- Review of the Animal Health & Welfare Delivery Landscape
- Responsibility and cost sharing for animal health and welfare
- GB Poultry register
- Bringing your pets to Britain
- Personal food & plant imports
- Consumer working

defra Department for Environment Food and Rural Affairs

Animal health and welfare

Contact us | Access to information | News | Site A-Z | Search

Animal Health & Welfare | Economics & Statistics | Environmental Protection | Exports & Trade | Farming | Marine & fisheries | Food & Drink | Horticulture | Plants & Seeds | Rural Affairs | Science | Sustainable Development | Wildlife & Countryside

Homepage > Animal health and welfare > **Home page**

Defra home page

Animal health and welfare
- Animal by-products
- Animal health & welfare strategy
- Animal welfare
- Bringing pets to the UK
- Disease surveillance & control
- International trade
- Livestock movements, ID & tracing
- Personal food & plant imports

Animal health and welfare

The health and welfare of animals is central to Defra's work of protecting and improving livestock and controlling and eradicating disease. The Animal Health and Welfare pages are divided into various subject areas each with its own 'Home page' and 'Latest news' section. These are detailed below.

Animal by-products
Information on the disposal of animal by-products.

Animal health & welfare strategy
Information relating to the strategy and the England Implementation Group.

Animal welfare
Latest information on the Animal Welfare Act and guidance on domestic and farm animal welfare.

Bringing pets to the UK
Guidance on the Pet Travel Scheme (PETS) and rabies quarantine.

Disease surveillance and control
Information on biosecurity, veterinary surveillance and notifiable diseases including:

- Avian influenza (Bird flu)
- Bluetongue
- BSE
- Equine Infectious Anaemia
- Foot and Mouth Disease
- Rabies
- Scrapie
- Tuberculosis

Looking for...
- News
- Publications
- Statistics
- Grants & funding
- Forms & licences
- Consultations

Quick links
- Animal Health Divisional Offices
- Animal health & welfare publications
- Animal Welfare Act 2006
- Avian Influenza (bird flu)
- Bluetongue
- Bringing pets to the UK
- Consumer working
- GB Poultry register
- Livestock markets roadshow

☒ GB Poultry Register

(c) Dissimilar versions

Figure 1. Similar and dissimilar versions of Web pages. The versions of (a) share the same information, they do not need to be crawled twice. The versions of (c) have the same banner and menus but the main information of the page is changed, a second crawling is then necessary.

without scrolling, (3) a machine learning based approach for supervised feature selection to increase prediction accuracy by eliminating noisy features.

2 Web page comparison

Two versions of a Web page are considered similar if the changes that occurred between them are not important enough to archive both of them. They are dissimilar otherwise (see Fig 1). To compare Web page versions, we first extract features from them as described below.

2.1 Visual descriptors

Important changes between page versions will often produce differences between the visual rendering of those versions. We propose to quantify these differences by computing and comparing the visual features in each page version. Each version is described as an image of its rendering capture (snapshot). We compute a visual signature on this captured image for each page. Images are first described by color descriptors, because they seem appropriate for Web page changes and are already used in Phishing Web page detection [9]. We also incorporate powerful edge-based descriptors with SIFT descriptors [10] because they give state-of-the-art performances in real image classification tasks.

For image representation, we follow the well-known Bag of Words (BoWs) representation [11, 16]. The vector representation of the rendered Web page is computed based on a sampling of local descriptors, coding and pooling over a visual dictionary. [7] Recent comparisons for image classification point out the outstanding performances of a regular dense sampling [2, 7]. We apply a first strategy called *whole Web page* feature, that samples regularly the visual representation of the whole page. However, the most significant information is certainly not equally distributed over the whole captured Web page. As noted in [17], the most important information is generally located in the visible part of pages without scrolling. A second strategy called *Top of Web page* feature, provides a visual vector using only the features located in the visible part of the page.

Since the visible part of a Web page depends on the size of the browser window, we take a generic window height of 1,000 pixels, greater than 90% of users' browser resolutions to ensure we do not miss information directly visible by most users. In the next sections, we will denote *the visible part of Web pages without scrolling*, also known as *above the fold*, by *top of Web pages*.

2.2 Structural descriptors

We extract various features directly from the code of Web pages. For instance, we extract Jaccard indices, a similarity value that indicates the preservation between versions of hyperlinks and of URL addresses of images. We assume that similar pages tend to keep the same hyperlinks and images.

We also extract some features from the difference tree returned by the VI-DIFF algorithm [14] that detects some operations between the VIPS structures of versions, e.g. insertions, deletions or updates of VIPS blocks, or even a boolean value returning whether two versions have the same VIPS structure. The more operations are detected, the less similar versions are assumed to be. We denote the features extracted from the VI-DIFF algorithm by VI-DIFF features.

2.3 Similarity between versions

Let V^A be the last archived version of a Web page and V^N the new version of the same Web page. We extract several visual and structural descriptors (see sections 2.1 and 2.2), and use different metrics (Euclidian, χ^2 distances, *etc*) to compare them. Heuristics may be used to set them individually and to select the best similarity function with a manually-tuned threshold to discriminate dissimilar pairs of Web pages from the similar ones.

We propose here an alternate scheme embedding all the similarity functions into a learning framework. Let the M visual feature/metric associations and the N structural similarities be aggregated in a vector **x** where:

$$\mathbf{x} = \begin{pmatrix} sim^1_{visual}(V^A, V^N) \\ \vdots \\ sim^M_{visual}(V^A, V^N) \\ sim^1_{struct}(V^A, V^N) \\ \vdots \\ sim^N_{struct}(V^A, V^N) \end{pmatrix} \qquad (1)$$

Combining both approaches seems appropriate to have a better understanding of the changes as perceived by human users. Learning combinations of complementary descriptors also makes the categorization task more efficient [15]. We investigate in the next section a statistical learning strategy based on a labeled dataset to classify the vectors **x**.

3 Supervised Feature Selection

There exist many families of distance learning methods [19]. In the context of image classification, Frome et al. [8] propose to learn a distance expressed as a linear combination of elementary distances: $D(\mathcal{F},\mathcal{I}) = \sum_j w_j^{\mathcal{F}} d_j^{\mathcal{F}}(\mathcal{I}) = \langle \mathbf{w}^{\mathcal{F}}, \mathbf{d}^{\mathcal{F}}(\mathcal{I})\rangle$ where $d_j^{\mathcal{F}}(\mathcal{I})$ are patch-to-image distances from a focal image $\mathcal{F}$ to a given image $\mathcal{I}$.

In our context, we are also interested in learning distances between versions in a supervised framework to determine whether two versions are similar or not. However, it is not a version classification problem as in [8]. Indeed, we do not want to classify samples (versions) but similarities. Moreover, our similarities are based on human judgement and allow subtleties as shown in Fig 1.

We then propose to express the learning of the combination of similarities as a binary classification in similarity space: for any couple of versions $(V^A, V^N)_i$, let their class $y_i = 1$ iff V^A and V^N are similar, -1 otherwise. Let $\mathbf{x}_i$ be a vector derived from heterogeneous similarities between V^A and V^N (as defined in subsection 2.3). We train a linear Support Vector Machine (SVM) to determine $\mathbf{w} = \sum_j \alpha_j y_j \mathbf{x}_j$ such that $\langle \mathbf{w}, \mathbf{x}_i\rangle = \sum_j \alpha_j y_j \langle \mathbf{x}_j, \mathbf{x}_i\rangle$ gives us the class of $(V^A, V^N)_i$. The similarity vectors $\mathbf{x}_j$ of training couples $(V^A, V^N)_j$ are used to train an SVM. For any test couple $(V^A, V^N)_i$, the trained SVM returns (1) whether $y_i = 1$ or $y_i = -1$, (2) whether V^A and V^N are similar or dissimilar, (3) whether V^N needs to be archived or not, with V^A already archived. Those three propositions are equivalent.

To study the contributions of the different types of features in the discrimination task, we first train a linear SVM with all the features. Each element w_k of **w** corresponds to the weight associated to the k-th similarity feature of **x**. Therefore, if the learnt w_k are close or equal to 0, the k-th similarity features of **x** are not determinant for categorization. Such similarities are considered noisy, irrelevant (not discriminant) in determining whether two versions are similar or not. To go one step further, we also propose a more explicit feature selection method based on the automatic *normal based feature selection* [12] that uses the fact that a feature k with the weight w_k close to 0 has a smaller effect on the prediction than features with large absolute values of w_k. Then features with small $|w_k|$ are good candidates for removal. The number of selected features may be set based on data storage and calculation constraints, or iteratively reduced using a validation set.

4 Experiments

A real corpus of data provided by the *Internet Memory Foundation*[1] is in our experiments. About 1000 pairs of Web pages captured from many different governmental Web sites from the United Kingdom about education, health, sport, ..., have been labelled as similar or not. The identical couples of versions are removed. Finally, 202 pairs of Web pages were extracted[2]: 147 and 55 (72.8% and 27.2%) couples of similar and dissimilar versions, respectively.

To compute visual similarities, we use SIFT [10] and color descriptors with visual codebooks of sizes 100 and 200. These are relatively small compared to the sizes used on large image databases but consistent with the size of our dataset. SIFT descriptors are calculated with the ColorDescriptor [18] software on a regular grid with a step size of 6 pixels. We use our own implementation for color desciptors extraction. We compute color histograms in the HSV space, on local patches of size 12x12 pixels. The BoWs of page versions are computed using the two strategies described in section 2.1: (1) over the rendering of whole Web pages and (2) the top of Web pages. Euclidian and χ^2 distances are then computed between the BoWs of successive page versions normalized using L^2-norm and L^1-norm, respectively. We also compute for each couple of page versions, the VIPS structures [5] and the VI-DIFF difference trees [14] from which we extract structural similarity values, e.g. the (symmetrized) ratio of identical nodes, boolean values on some criteria such as an identical VIPS structure (Fig 2.b). In the end, we have 16 visual and 25 structural features.

4.1 Similarities

We first report some similarity distributions for similar and dissimilar Web page couples in Fig 2. For visual (Fig 2.a) and structural (Fig 2.b,c) similarities, similar samples tend to be close to each other, and dissimilar samples tend to be more distant, even with basic distances. In Fig 2.d, structural and visual distances are drawn together. We observe the same behaviour for all our similarity measures: none of them allows to clearly discriminate similar from dissimilar samples. Instead of using them individually, we propose to combine those different similarities in a binary classification scheme that returns whether a couple of versions are similar or not by using their vector of similarity scores.

4.2 Binary classification

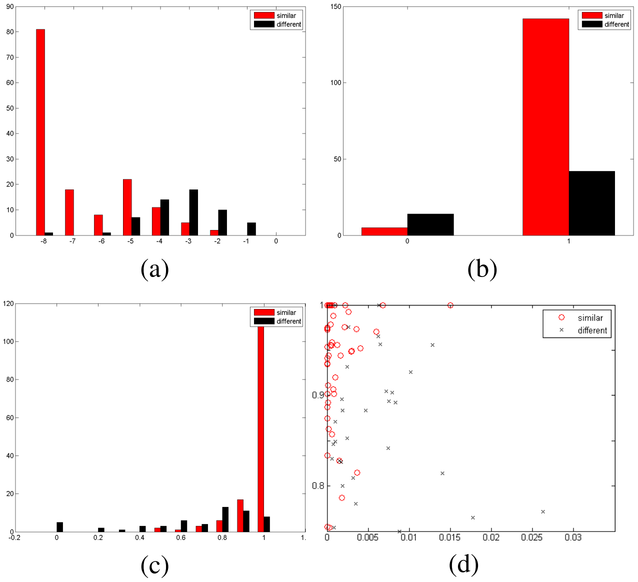


Figure 2. Distributions of Web page distances or similarities: (a) decimal orders of magnitude of visual χ_2-distances using top of Web page SIFT features with 200 codewords, (b) boolean similarity of VIPS tree structures between two versions, (c) Jaccard indices of hyperlinks, (d) Jaccard indices of hyperlinks as a function of visual χ_2-distances using top of Web page SIFT features with 100 codewords. The similar and dissimilar couples of versions are in red (circles) and black (crosses), respectively.

We use leave-one-out cross-validation to evaluate the performance of our model. We compare our results to the random classifier which automatically predicts the most represented class in the dataset, yielding a baseline accuracy of 72.8%.

4.2.1 Evaluation of visual features

We first use only the visual information of pages. Structural variables of **x** are ignored. The prediction accuracies when selecting different subsets of local descriptors (SIFT and color) sampled on whole Web pages or top of Web pages are presented in Table 1. SIFT and color descriptors achieve good performances for Web page change detection. Using the visible part of Web pages without scrolling (87.1%) is also a lot more discriminant than using whole Web pages (83.2%). Combining both of them gives even worse results (85.1%) than using only the top of Web pages (87.1%). Important changes are more likely to be directly observable whereas changes at the bottom of Web pages, often advertisements, are more likely to be less important and noisy. The accuracies obtained validate our approach.

[1] http://internetmemory.org/

[2] The corpus we used in this work can be provided on demand.

Selected Visual Features		Accuracy (%)
Whole Web page	Top of Web page	
None	SIFT	84.2
None	color	82.7
None	SIFT + color	**87.1**
SIFT	None	79.7
color	None	80.7
SIFT + color	None	83.2
SIFT + color	SIFT + color	85.1

Table 1. Prediction results when using visual features.

Selected Structural Features		Accuracy (%)
Jaccard Indices	VI-DIFF	
Yes	No	85.1
No	Yes	76.7
Yes	Yes	**87.6**

Table 2. Prediction results when using structural features.

4.2.2 Evaluation of structural features

We study in Table 2 the prediction accuracies when different subsets of structural similarities only are used. Jaccard Indices of links are the most discriminant structural features (85.1%) but the other structural features extracted from VI-DIFF are still informative.

4.2.3 Structural and visual feature combination evaluation

We investigate the combination of structural and visual features in Table 3. The accuracy when combining all of them (90.1%) is better than when using only structural (87.6%) or visual (87.1%) features. Visual and structural features are then complementary.

Furthermore, we propose to combine in Table 3 the visual and structural features that gave the best accuracies in previous sections. An exhaustive manual selection among all the 41 structural and visual features to find the set that maximizes prediction would be too time-consuming. The accuracy is improved up to 93.1% when combining only Jaccard indices of links and the top of page visual representations. But the best choice in the context of archiving is to minimize the number of pages actually dissimilar but predicted similar. This happens when we select all the structural features and the visual features of the top of pages. We then obtain an overall accuracy of 92.1% comparable to the best accuracy obtained manually (93.1%), but only 12.7% of dissimilar versions predicted similar instead of 14.5%.

Selected Feature similarities		Acc. (%)
Structural	Visual	
All	All	90.1
All	Top of Web page	92.1
Jaccard indices	All	91.6
Jaccard indices	**Top of Web page**	**93.1**

Table 3. Prediction results when using structural and visual features.

Selected features	1	4	8	10	13
Accuracy (%)	84.7	88.1	91.1	91.6	92.6
Selected features	15	18	22	27	41
Accuracy (%)	92.6	91.1	90.6	90.1	90.1

Table 4. Accuracy as a function of the number of selected features with the normal based feature selection (highest absolute values).

We also investigate the automatic *normal based feature selection* method described in section 3 in Table 4. The best accuracy obtained with that automatic method is 92.6% when the 13 to 15 features with the highest absolute values in **w** are selected. It is comparable to our best accuracy of 93.1% (Table 3) with 10 features selected.

5 Scalability Issue

We are involved in the EU SCAPE project whose goal is scalable digital data preservation. The scalability of Web archiving technologies is hence decisive. Our first requirement is to perform the comparison of two page versions in about one second on a standard PC with single CPU.

	Whole Page	Top of Page	**Total time (s)**
SIFT	7.25	5.06	12.31
color	0.61	0.38	0.99
SIFT BoWs	2.54	1.78	4.32
color BoWs	0.43	0.24	0.67
Distances			0.28
SVM			0.04
Total time (seconds)	**11.15**	**7.78**	**18.61**
Accuracy (%)	**93.1**		

Table 5. Running time baseline (in seconds).

To evaluate our system, we compute on a single 3.47GHz PC the average running times of comparison on our dataset

(Table 5). The captured images of Web pages have an average width and height of 929 (±251) and 1272 (±723) pixels, respectively. The brute force time of the whole process is close to 20 seconds. The bottleneck is the SIFT computation and the assignment process of BoWs. We use ColorDescriptor [18] to compute SIFT, which is quite slow compared to other recent packages now available. Some other descriptors like SURF [3] and CARD [1] could be tested to speed up recognition, but their impact on recognition accuracy is yet to be determined.

Another way to speed up computation concerns the method setting. First, we only consider the Top of Web page component, second, we analyze the lost in accuracy as the step size parameter increases. Table 6 shows the time and accuracy performances. Both descriptor combinations are tested. When using a large step, the computation time significantly decreases while keeping a correct detection accuracy. The prediction can be processed in less than two seconds with an accuracy of 88%. Those results allow us to respect the SCAPE Project requirements.

Step (pixels)	6	12	24	32
Time (seconds) (SIFT)	7.16	2.33	1.45	1.29
Accuracy (%) (SIFT)	88.1	87.6	86.6	84.6
Time (seconds) (SIFT+color)	7.78	2.65	1.77	1.61
Accuracy (%) (SIFT+color)	93.1	88.1	88.1	85.6

Table 6. Influence of the step parameter on time and accuracy.

6 Conclusion

In this paper, a complete Web page comparison framework for Web archiving is presented. Both structural and visual features are used to analyze pages of Web sites, and estimate when or with which frequency they must be visited. We introduced a scheme to learn an optimal similarity combination. Many features and similarities have been tested. One of the main results is that important changes generally appear at the visible part of pages without scrolling.

Experiments on real data have been presented. A large set of couples of pages manually labelled by archivists has been used for a quality evaluation of our visual and structural similarity method. They confirm that both structural and visual informations are useful for change detection. Web archiving and digital preservation are major challenges for our digital societies. We are convinced that the Multimedia Processing community will bring important support and technical solutions to many of their issues.

Acknowledgments This work was partially supported by the SCAPE Project. The SCAPE Project is co-funded by the European Union under FP7 ICT-2009.4.1.

References

[1] M. Ambai and Y. Yoshida. Card: Compact and real-time descriptors. *ICCV*, 2011.

[2] S. Avila, N. Thome, M. Cord, E. Valle, and A. Araújo. Bossa: Extended bow formalism for image classification. In *ICIP 2011*.

[3] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. *ECCV*, 2006.

[4] M. Ben Saad, S. Gançarski, and Z. Pehlivan. A novel web archiving approach based on visual pages analysis. In *IWAW 2009*.

[5] D. Cai, S. Yu, J. Wen, and W. Ma. Vips: a vision-based page segmentation algorithm. *Microsoft Technical Report, MSR-TR-2003-79-2003*, 2003.

[6] J. Cao, B. Mao, and J. Luo. A segmentation method for web page analysis using shrinking and dividing. *International Journal of Parallel, Emergent and Distributed Systems*, 25, 2010.

[7] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. *BMVC*, 2011.

[8] A. Frome, Y. Singer, and J. Malik. Image retrieval and classification using local distance functions. In *NIPS 2006*.

[9] A. Fu, L. Wenyin, and X. Deng. Detecting phishing web pages with visual similarity assessment based on earth mover's distance (emd). *IEEE TDSC*, 3, 2006.

[10] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60, 2004.

[11] W. Ma and B. Manjunath. Netra: A toolbox for navigating large image databases. In *ICIP 1997*.

[12] D. Mladenić, J. Brank, M. Grobelnik, and N. Milic-Frayling. Feature selection using linear classifier weights: interaction with classification models. In *SIGIR 2004*.

[13] M. Oita and P. Senellart. Deriving dynamics of web pages: A survey. In *Temporal Workshop on Web Archiving, Hyderabab (India)*, March 2011.

[14] Z. Pehlivan, M. Ben Saad, and S. Gançarski. Vi-DIFF: Understanding Web Pages Changes. In *DEXA 2010*.

[15] D. Picard, N. Thome, and M. Cord. An efficient system for combining complementary kernels in complex visual categorization tasks. In *ICIP 2010*.

[16] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV 2003*.

[17] R. Song, H. Liu, J. Wen, and W. Ma. Learning block importance models for web pages. In *WWW 2004*.

[18] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *PAMI*, 32(9):1582–1596, 2010.

[19] L. Yang and R. Jin. Distance metric learning: A comprehensive survey. *Michigan State University*, pages 1–51, 2006.