Browsing Documents on a Dense Embedding that Captures Theme Evolution

Alessandro Perina
Microsoft Research
One Microsoft Way, Redmond, WA
alperina@microsoft.com

Nebojsa Jojic
Microsoft Research
One Microsoft Way, Redmond, WA
jojic@microsoft.com

ABSTRACT

We describe a new interaction strategy for browsing documents consisting of text and images. The browser represents a collection of documents as a grid of key words with varying font sizes that indicate the words' weights. The grid is computed using the counting grid model [7], so that each document approximately matches in its word usage the word weight distribution in some window (6 × 6 in our experiments) in the grid. In comparison to other document embedding approaches, this strategy leads to denser packing of documents and higher relatedness of nearby documents: The two documents that map to overlapping windows literally share the words found in the overlap. This leads to smooth thematic shifts that can provide connections among distant topics on the grid. The images are embedded into the appropriate locations in the grid, so that a mouse over any location can invoke a pop-up of the images mapped nearby. Once the user locks on an interesting spot in the grid, the summaries of the actual documents that mapped in the vicinity are listed for selection. In this document browser the arrangement of related words and themes on the grid naturally guides the user's attention to topics of interest. For an illustration we describe and demonstrate (in video submission) a browser of four months of CNN news.

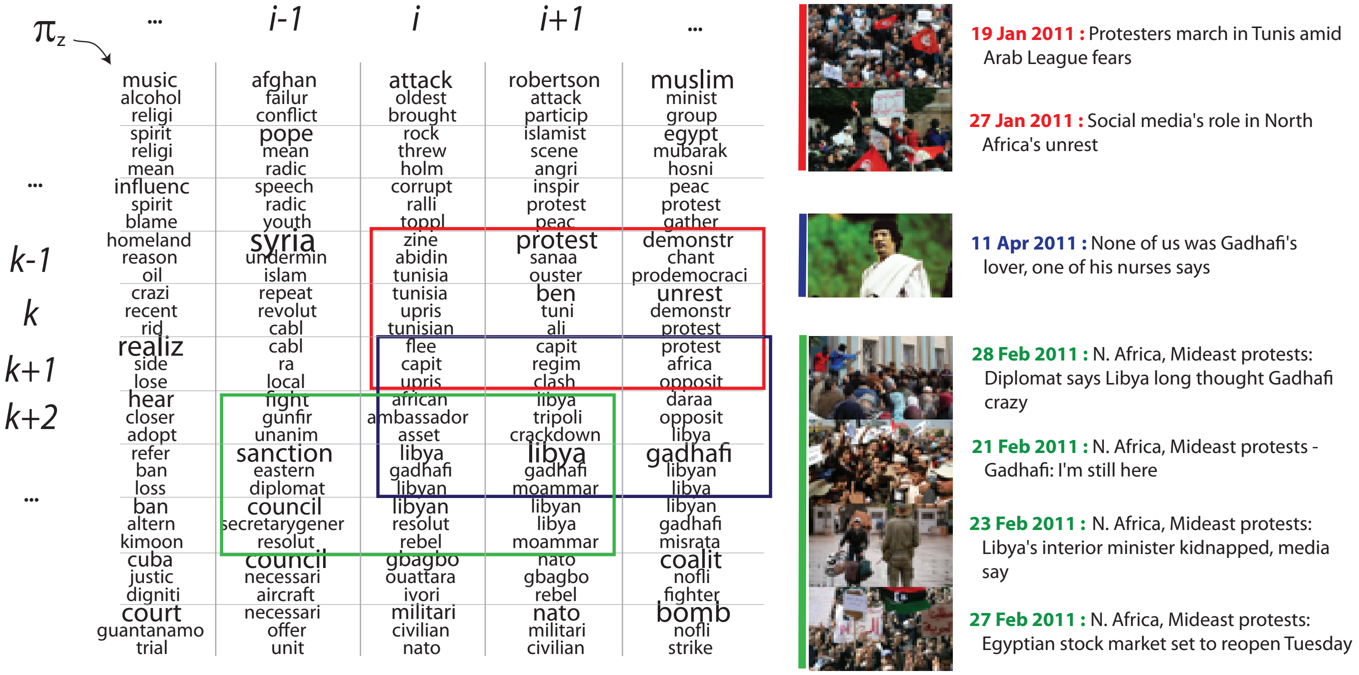
INTRODUCTION

Summarizing, visualizing and browsing text corpora are important problems in computer-human interaction. As the data becomes more massive, ambiguous, or conflicting, it may become hard for people to glean insights from it. To help the users, researchers have developed several visual analytics tools facilitating the analysis of such corpora. These tools are used to interactively make sense of complex datasets, a process referred to as *sensemaking* [10].

We describe a new approach to browsing documents consisting of text and images, e.g. news stories on the web, social media, special interest web sites, etc. The browsing through documents is based on the exploration of the hidden variable of the on the counting grid (CG) generative model [7], which has recently been used for a variety of tasks related to regression and classification. The counting grid model represents the space of possible documents as a grid of word counts. Each individual document is mapped to a window into this grid so that the tally of these counts approximately matches the word counts in the document. The grid can vary in size, and so can the window. As the documents are allowed to be mapped with overlap, in order to maximize the likelihood of the data, the learning algorithm has to map similar documents to nearby locations in the grid, so that the words that the two documents share appear in the grid positions in the overlap of the corresponding windows. This leads to a compact representation where the theme of the documents smoothly varies across the grid, achieving a higher density of packing than previous embedding approaches (e.g. Egypt unrest news are placed close to other stories about Arab Spring, with Libya taking another distinct location in that area of the CG; nearby are stories about oil prices, and near these are more stories about the markets and economy, near which are stories referring to Fed's Bernanke, near which are stories about congress and the President, which, in a counting grid defined on a torus may loop back to Libya through military themes.) To provide natural means of summarization and browsing of the documents, we render a CG representation based only on the most frequent words in each position. We further embed the images from each document into the appropriate locations in the counting grid, so that they can pop up when the user focuses on a particular area of the grid (e.g. by mouse over). This provides the user with both a global and local perspective on the underlying set of documents and their relationships, without observing directly the underlying documents, but rather the CG model's representation of the document space. Once the user locks on an interesting spot in the grid, the summaries of the actual documents that mapped in the vicinity are listed for selection. This idea leads to an intuitive document browser that is especially well suited to touch devices, where moving a cursor is the most natural interaction modality, while typing is particularly difficult. Additionally, the interface assists the user in discovering documents of interest without having to define a particular target and associated keywords first: The arrangement of related words and themes on the grid naturally guides the user's attention to topics of interest. For an illustration we describe and demonstrate (video submission) a browser of four months of CNN news from winter and spring 2011, a period particularly rich in news-worthy events.

COUNTING GRIDS (CGS)

CHI'13, April 27–May 2, 2013, Paris, France.



π_z	...	i-1	i	i+1	...
	music alcohol religi	afghan failur conflict	attack oldest brought	robertson attack particip	muslim minist group
	spirit religi mean	pope mean radic	rock threw holm	islamist scene angri	egypt mubarak hosni
...	influenc spirit blame	speech radic youth	corrupt ralli toppl	inspir protest peac	peac protest gather
k-1	homeland reason oil	syria undermin islam	zine abidin tunisia	protest sanaa ouster	demonstr chant prodemocraci
k	crazi recent rid	repeat revolut cabl	tunisia upris tunisian	ben tuni ali	unrest demonstr protest
k+1	realiz side lose	cabl ra local	flee capit upris	capit regim clash	protest africa opposit
k+2	hear closer adopt	fight gunfir unanim	african ambassador asset	libya tripoli crackdown	daraa opposit libya
...	refer ban loss	sanction eastern diplomat	libya gadhafi libyan	libya gadhafi moammar	gadhafi libyan libya
	ban altern kimoon	council secretarygener resolut	libyan resolut rebel	libyan libya moammar	libyan gadhafi misrata
	cuba justic digniti	council necessari aircraft	gbagbo ouattara ivori	nato gbagbo rebel	coalit nofli fighter
	court guantanamo trial	necessari offer unit	militari civilian nato	nato militari civilian	bomb nofli strike

19 Jan 2011 : Protesters march in Tunis amid Arab League fears

27 Jan 2011 : Social media's role in North Africa's unrest

11 Apr 2011 : None of us was Gadhafi's lover, one of his nurses says

28 Feb 2011 : N. Africa, Mideast protests: Diplomat says Libya long thought Gadhafi crazy

21 Feb 2011 : N. Africa, Mideast protests - Gadhafi: I'm still here

23 Feb 2011 : N. Africa, Mideast protests: Libya's interior minister kidnapped, media say

27 Feb 2011 : N. Africa, Mideast protests: Egyptian stock market set to reopen Tuesday

Figure 1. A part of the counting grid trained on the news stories. Three windows are highlighted along with seven stories that mapped there. Color indicates the mapping. The movement through the grid captures the spread of the Arab Spring in North Africa, and the subsequent UN reaction.

The counting grid consists of a set of discrete locations indexed by ℓ in a map of arbitrary dimensions (30×30 to 40×40 2D torus grids in examples here). A part of a counting grid is illustrated in Fig 1. Each location contains a different set of weights for the Z words in the vocabulary ($Z = 10000$ here). The weight of the z-th word at location ℓ is denoted by $\pi_{z,\ell}$ and the weights add up to one, $\sum_\ell \pi_{z,\ell} = 1$. Thus π is a probability distribution over words and defines the local word usage proportions. (These weights are partially illustrated in Fig. 1 using font size variation, but showing only the top 3 words at each location.) A document has its own word usage counts c_z and the assumption of the counting grid model is that this word usage pattern is well represented at some location k in the grid in the following way: When a window of a certain size is placed at location k in the CG, and the CG weights are averaged across N CG locations in the window W_k to obtain $h_z = \frac{1}{N}\sum_{\ell\in W_k}\pi_{z,\ell}$, then this distribution is approximately proportional to the observed document counts $h_z \propto c_z$. In other words, approximately the same words in the same proportions are used in the document and in its corresponding counting grid window W_k. The window size 6×6, and thus $N = 36$ was used in our experiments, but due to space limitations 3×3 windows were used in Fig. 1.

The CG estimation algorithm [7] starts with a random initialization which gives all words roughly equal weights everywhere. The subsequent iterations (re)map the documents to the windows in the grid and rearrange words to match the weights currently seen in the grid. In each iteration, after the mapping, the grid weights at each location are re-estimated to match the counts of the mapped document words. We found that the algorithm converged in 70-80 iterations, which sums up to minutes for summarizing months of news on a single standard PC. As this EM algorithm is prone to local minima, the final grid will depend on the random initialization, and the neighborhood relationships for mapped documents may change from one run of the EM to the next. However, as shown in the supp. material, the grids qualitatively always appeared very similar, and some of the more salient similarity relationships were captured by all the runs (e.g. the Arab Spring news that referred to multiple different countries with very different unfolding of events are always grouped nearby). More importantly, a majority of the neighborhood relationships make sense from a human perspective and thus the mapping gels the documents together into logical, slowly evolving themes. As discussed below, this helps guide our visual attention to the subject of interest. As the algorithm optimizes the likelihood of the data, all resources (grid locations) must be used, and the packing is much denser than in the previous embedding approaches, thus occasionally squishing themes together even though no documents map to their interface. In our opinion, it is a small price to pay for high real estate utilization and, for the most part, intuitive arrangement of themes

MULTIMODAL CG DISPLAY AND BROWSING

To browse a collection of multimodal documents consisting of both text and images, we first fit a CG model to the corpus, and then embed the images into appropriate locations of the grid, so that each image is placed in the grid position in the center of the window to which the source document was mapped (Fig. 2). This results in a grid of images of the same size as the word counting grid with a rough semantic alignment: In each image's vicinity the grid locations have high weights on the words related to the image. Obviously, there is now a multitude of possible approaches to visualizing this embedding in a way that explores the two modalities in concert. To show the image embedding, we can simply show a tiling of images (e.g. based on the 30 × 30 CG). In locations where multiple images are mapped, we can pick one at random (as in our experiments), or the one that was used in multiple documents, or the one selected by a computer vision algorithm. In addition, the images mapped to the same location can slowly cycle. To visualize the CG word weights $\pi_{z,\ell}$ in each grid location, we show the top k words ($k = 3$ in our experiments) using the font size to indicate the word weight. In our browser, we can switch between the two representations, or show them one on top of the other with a certain level of transparency (Fig. 2). In addition, a pointer (a mouse cursor, fingertip on touch devices, etc.) can be used to force the switch between images and words locally in a window of a certain size (5 × 5 in our experiments). In this way the user can base their exploration primarily on one modality, bringing the other modality to the fore by hovering over the grid parts of interest. In particular, we find the word representation particularly useful in drawing the user's attention across related themes to the point of interest. As the user naturally moves the pointer toward their eyes' focal point the pointer uncovers images underneath to further refine the user's understanding of the grid content. At any point, the user can stop and indicate (e.g. by a click) their desire to see the source documents that mapped in this region. We implemented two ways of uncovering the images in the region where the user hovers. In the first approach, the words in the grid locations around the cursor are highlighted and the images from these locations are shown next to the highlighted area. In the second approach, we simply replace the area around the cursor with images. As the embedding is based on overlapping windows, in both cases it is possible that some of the images that pop up this way are related to the themes slightly outside the highlighted area. Once the user is used to this it becomes imperceptible as the matching words (or images) are never far and slight movements of the pointer help lock onto the topic of interest. To further indicate the smooth nature of the mapping, we experimented with varying sizes and intensities of images that pop up. For example, in Fig. 2 the central image of the highlight is of larger size and it slightly overlaps the 6 images around it, which themselves are larger and overlap even more the images around them, creating an impression of the underlying images popping out from the words, with relationship being approximate but smooth, inviting the user to move the cursor around.

Although the CG model glues the documents together based on the vocabulary overlap that can contain a large number of different words, to a human observer, just the top words for each location seem to provide enough insight into the thematic shifts in the grid. The grid in Fig. 2 gels the disaster stories together due to their common vocabulary (e.g. disaster, response, emergency, etc.), but in the browser most of that shared vocabulary is overtaken by the words that get high weight in individual locations (earthquake, tornado, airplane, crash, snow, storm, etc.). The human mind easily detects connections among these and need not observe all of the "glue" that linked these topics together. In our experience, the CG visualization seems to stimulate the user's own associations and memory and guides the user to the target even if they did not start with a particular target in mind: A look at a salient Japan and earthquake keywords creates an association with local weather disasters, reminding the user that they were following an airplane crash story. This association process is guided by CG's own 'associations' so that the spot in the grid is found quickly. Further interaction with the grid to invoke visual stimulus increases the pace of news discovery.

To accommodate for variable display sizes and corpora diversities, we can train a hierarchy of CG models of various sizes, where model of one size is initialized by an upsampled version of the model of the smaller size. In this multi-granular approach, the user can zoom in and out of any part of the grid. Window size choice provides the tradeoff between finer document overlaps and the computational complexity of the CG estimation, but for the CNN news stories at least, the latter was not a limiting factor.

DISCUSSION

Our approach provides some important advantages over the existing visualization/browsing/search approaches. The 10x10 grid website [1] also arranges images into a grid. But, the placement of images is not optimized so that the nearby locations capture related stories. Previous methods for spatially embedding documents [5, 2] produce sparse representations (e.g. "The Galaxy of News" [8]), which are only locally browsable, whereas the counting grids use the screen real estate much more efficiently. In addition, our approach allows embedding of multiple modalities. Various galaxy approaches required that the user interact with the embedding through the statistical model, manipulating its parameters and/or weights, which may be impenetrable to the user, thus requiring a laborious guess and check strategy [1, 6]. This issue is still a subject of research in HCI [3]. In contrast, the CG parameters (grid size and the scope of overlap, i.e. the window size), are more intuitive, and multi-granular approaches may remove a need for parameter selection altogether.

The CG visualization reminds one of *tag clouds*, visual representations that indicate frequency of word usage within textual content. Google News Cloud [2] sorts words alphabetically, varying the font based on the relevance. If a word is selected other similar words are highlighted. But the links among the complex documents that combine a variety of words are not evident. Other tools (e.g., Toronto Sun, Washington Post websites) cluster words based on co-occurrence or proximity and then position the words belonging to the same clusters near each other and use color to emphasize the structure. Still, the words are not spatially embedded within a cluster, and so only cluster hopping can be performed, in contrast with smooth thematic drifts found in CGs. For the most part, the tag clouds are designed to provide a useful and visually pleasing summary of the news [9, 4], rather than a

[1] `http://www.tenbyten.org/10x10.html`

[2] `http://fserb.com.br/newscloud/index.html`

two-dimensional densely organized multimodal browsing index which CG provides. In terms of providing a means for traversing an organization of news, our method shares some similarities with *Newsmaps*[3] which use a hierarchical representation, a tree. But the traversal paths descend along the branches of the tree while CGs often capture many different directions of thematic drifts which can loop back.

REFERENCES

1. Alsakran, J., Chen, Y., Zhao, Y., Yang, J., and Luo, D. STREAMIT: Dynamic visualization and interactive exploration of text streams. 131–138.
2. Chen, Y., Wang, L., Dong, M., and Hua, J. Exemplar-based visualization of large document corpus (infovis2009-1115). *IEEE Transactions on Visualization and Computer Graphics 15* (2009), 1161–1168.
3. Endert, A., Fiaux, P., and North, C. Semantic interaction for visual text analytics. In *ACM CHI* (2012), 473–482.
4. Helic, D., Trattner, C., Strohmaier, M., and Andrews, K. Are tag clouds useful for navigation? a network-theoretic analysis. *Journal of Social Computing and CyberPhysical Systems 1* (2011), 33–55.
5. Iwata, T., Yamada, T., and Ueda, N. Probabilistic latent semantic visualization: topic model for visualizing documents. In *ACM KDD* (2008), 363–371.
6. Jeong, D. H., Ziemkiewicz, C., Fisher, B. D., Ribarsky, W., and Chang, R. ipca: An interactive system for pca-based visual analytics. *Comput. Graph. Forum 28* (2009), 767–774.
7. Jojic, N., and Perina, A. Multidimensional counting grids: Inferring word order from disordered bags of words. In *UAI* (2011), 547–556.
8. Rennison, E. Galaxy of news: An approach to visualizing and understanding expansive news landscapes. In *ACM Symposium on User Interface Software and Technology* (1994).
9. Sinclair, J., and Cardew-Hall, M. The folksonomy tag cloud: when is it useful? *J. Inf. Sci. 34* (2008), 15–29.
10. Thomas, J., and Cook, K. *Illuminating the Path: The Research and Development Agenda for Visual Analytics.* IEEE Press, 2005.

[3] `http://newsmap.jp/`

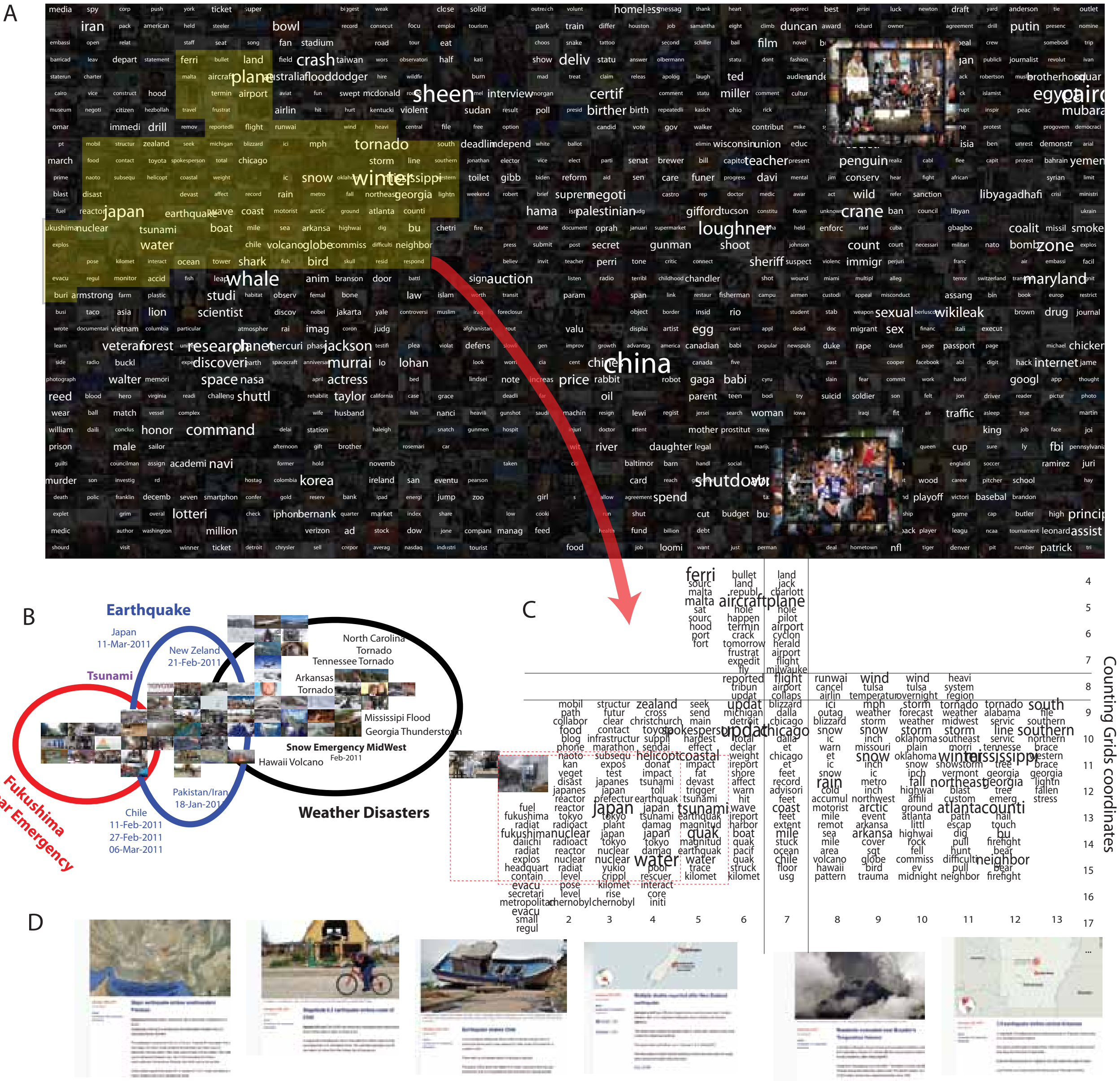


Figure 2. Browsable counting grid. A. The text and image representation of the grid are combined with emphasis on text. In two locations images are brought into the foreground. The grid is defined on a torus (with left matching the right and the top continuing at the bottom). Various theme drifts are visible, e.g. the japan-tsunami-water-whale-study- scientist-research-development-space-shuttle-nasa-command-navy semicircle on the left, or the region emphasized in B) and C) which captures the various disasters from the period. More can be seen in figures in supplemental material and the video submission. The preprocessing of the words reduced them to their roots and also made other standard alterations used in text analysis, but the unaltered words can be shown instead. B. Images mapped in the highlighted area. C. More of the top words in the highlighted area, and an illustration of how the images were embedded: As each document maps onto a window, the images from the document go to a location in the window (top left in the illustration to avoid clutter, but the middle of the window in actual implementation to provide more natural alignment). D. Some of the news that mapped to the highlighted area. The area of interest can be selected by cursor hover and the news can be recalled by a simple click.