Towards a multi-modal Deep Learning Architecture for User Modeling

Ange Tato[1], **Roger Nkambou**[2]
[1]École de Technologie Supérieure (ETS)
1111 Notre-Dame W. st., Montréal, Quebec, H3C 6M8, Canada
[2]Université du Québec à Montréal (UQAM)
201 President-Kennedy Avenue, Montreal, Quebec, H2X 3Y7, Canada
ange-adrienne.nyamen-tato@etsmtl.ca, nkambou.roger@uqam.ca

Abstract

Deep learning has succeeded in various applications, including image classification and feature learning. However, there needs to be more research on its use in Intelligent Tutoring Systems or Serious Games, particularly in modeling user behavior during learning or gaming sessions using multi-modal data. Creating an effective user model is crucial for developing a highly adaptive system. To achieve this, it is necessary to consider all available data sources to inform the user's current state. This study proposes a user-sensitive deep multi-modal architecture that leverages deep learning and user data to extract a rich latent representation of the user. The architecture combines a Long Short-Term Memory, a Convolutional Neural Network, and multiple Deep Neural Networks to handle the multi-modality of data. The resulting model was evaluated on a public multi-modal dataset, achieving better results than state-of-the-art algorithms for a similar task: opinion polarity detection. These findings suggest that the latent representation learned from the data is useful in discriminating behaviors. This proposed solution can be applied in various contexts where user modeling using multi-modal data is critical for improving the user experience.

Introduction and Related Work

User modeling is the process of using observable information, such as user actions or utterances, to infer unobservable information about a user (Zukerman and Albrecht 2001). As stated in (Birk et al. 2015), a user model that accurately represents a user over time in an Intelligent Tutoring System (ITS) or serious game leads to efficient adaptation, which increases learner/player satisfaction and motivation. To this end, it is important to ensure the effectiveness of the learner model before deploying the system for real use. With recent advances in Artificial Intelligence, why not think of an up-to-date user model? One that will be able to exploit deep learning techniques and multi-modal user data to describe the learner better and thus improve its learning gain. Multimodal learning involves relating information from multiple sources (Ngiam et al. 2011) (Jing et al. 2020). We propose a User-Sensitive Deep multi-modal Network Model (Us-DMN), which can be seen as a standalone user model or plugged into an existing standard user model to enrich its features. This model is intended to augment a user model with the ability to detect which kind of behavior the user exhibits automatically. The model is inspired by well-known deep-learning multi-modal architectures (Dimitri 2022).

ITS and Serious Games are learning environments that keep track of user data from different sources, which makes the proposed model useful for those systems. It can take advantage of all the data gathered from all the sources. The user data given as input to the deep learning architecture help improve the ability to predict user behaviors well. Some works tend to hand-code user features, based on experts' theory (Nkambou et al. 2016) while others explicitly define rules on how different behaviors can appear (Ramesh et al. 2013). In those conditions, Bayesian networks are sometimes used to represent the current user state (Tato et al. 2017a).

Multimodality is inherent to human behavior in general. Therefore, it is important to consider all sources of information about the user's current state to design an effective representation based only on data. Moreover, some key features of the user's behavior cannot be ignored since they can influence the interpretation of the collected data.

Deep learning techniques have been used for user modeling, especially for knowledge tracing. Deep Knowledge Tracing (DKT) (Piech et al. 2015) is the best example of a deep learning architecture that aims to model how students' knowledge evolves during learning. DKT and its multiple versions (Liu et al. 2019) (Song et al. 2021) (Song et al. 2022) can capture the sequential aspect of data and predict student performance based on the pattern of their responses sequentially over time. However, DKT is designed for a single modality: students' answers (often textual). Also, some research model the user in recommendation systems such as Elkahky (Elkahky, Song, and He 2015), which uses a deep neural network in an online service for automatic personalization to recommend relevant content to a large number of users. Their model matches rich user features to item features of service to learn relevant user behavior patterns and provide more adapted recommendations. Also, the model proposed by Covington (Covington, Adams, and Sargin 2016) for video recommendations is quite similar to ours, but they concatenate the user data altogether before extracting relevant features for the recommendation. Other re-

search uses convolutional neural networks (CNN) to extract user intentions from Medical Queries (Zhang et al. 2016). Duyu (Tang et al. 2014) has created a UCWCVM, a user-word composition vector model to effectively incorporate user information in the neural network for review rating prediction. Although our goal is similar to those solutions, they do not take advantage of the inherent multi-modality of user behaviors as the Us-DMN model does. Our solution can replace any of those models for the same task, but the reverse is not necessarily true.

Concerning the model architecture, "Each modality is characterized by very distinct statistical properties that make it difficult to ignore that they come from different input channels" (Srivastava and Salakhutdinov 2012). High-level features can be learned from multi-modal data by merging the modalities into a joint representation that captures the specificities to which the data correspond. There are, nevertheless, some methods, techniques, and tools for learning features from multi-modal data (affective data analysis, image classification). Unfortunately, many of these techniques are trying to learn those features separately for each modality (decision level approach (Poria et al. 2015)) and combining them by concatenation, averaging, or weighted voting (Zeng et al. 2007) (Gunes and Piccardi 2007). The risk of doing this in our case is that: 1) because the characteristics of a user are not independent, key correlation and significant information between features extracted from different modalities (Text, facial expression) can be missed; 2) Runtime and response time may be slow (especially if the analysis must be done in real-time) as each modality will have to be considered one at a time. To address these issues, we opt for a feature-level approach whose purpose is to combine the characteristics extracted from each input channel into a "joint vector" that we will call the user features vector (the penultimate layer of the proposed architecture). Most existing techniques for extracting latent characteristics operate in an unsupervised manner, such as the well-known Deep Boltzmann Machines (Wang et al. 2016). Since our problem is reduced to a supervised classification problem, these techniques will be reserved for contexts without labeled data.

Our model is the first multi-modal deep neural network model designed for user modeling that integrates multiple data from multiple sources and multiple deep learning models simultaneously and in a parallel fashion. More specifically, most existing data-driven approaches for user modeling are developed based on a single deep model and can hardly maintain good generalization across various data. Multiple deep learning architectures for multi-modal data have been proposed, such as the hybrid deep neural network (HDNN) model for Remaining Useful Life (RUL) estimation proposed by Ali et al. (Al-Dulaimi et al. 2019) which consists of two parallel paths (one LSTM and one CNN) followed by a fully connected multilayer fusion neural network, which acts as the fusion center that combines the two paths' output to form the target RUL. Huang et al. (Huang et al. 2020) integrate both clinical and imaging data using a multi-modal fusion model architecture capable of utilizing both pixel data from volumetric Computed Tomography Pulmonary Angiography scans and clinical patient data from the EMR to classify Pulmonary Embolism (PE) cases automatically. Multiple multi-modal architectures exist in many contexts but very few in the educational domain.

The Us-DMN can extract high-level features from low-level human-centered data by considering the multiple user interaction modalities. Konrad et al. (Gadzicki, Khamsehashari, and Zetzsche 2020) show a clear performance improvement with a multimodal fusion rather than a unimodal approach. The model has been tested on a multimodal dataset of debates where Emotional states, EEG (electroencephalography) data, and Textual opinions of participants were captured. We access the Us-DMN on its ability to extract a meaningful representation of a learner by automatically detecting the polarity of opinions based on all the data extracted. Our model is adaptable depending on the context because the penultimate layer is the extracted user features. The last layer can be replaced by any classes one might be willing to detect from user data (such as behaviors types, people whose game the system or not, etc.). The Us-DMN is built from three deep learning models: The Convolutional Neural Network (CNN) for its ability to extract useful features from data, the Long Sort-Term Memory (LSTM) for its ability in temporal modeling and the Deep Neural Network (DNN) for mapping features to a more separable space. Our model is slightly similar to that of Tara (Sainath et al. 2015) (they combined a Convolutional, Long Short-Term Memory and Fully Connected Deep Neural Networks). However, their model is intended for voice search, while our solution is multimodal and aims to understand better and enrich the user model. Also, our model is multimodal.

In the following sections, we present the proposed user-sensitive deep multimodal architecture, followed by the experiment's setup, the results, and discussions of the performance assessment of the model. The paper ends with some concluding remarks and future works.

The User-sensitive Deep multi-modal Network (Us-DMN) Architecture

We opted to use feedforward DNNs for non-sequential modalities due to their ability to extract meaningful information by delving deep into data. DNNs are neural networks consisting of multiple layers, with "Deep" indicating networks with more than one hidden layer. In DNNs, each layer of nodes is trained on a unique set of features based on the previous layer's output. As you progress through the neural net, the nodes can recognize increasingly complex features as they aggregate and recombine features from the previous layer. Since our proposed model is intended for real-time use, we took execution time into account as a crucial factor. Therefore, we limited the number of layers in the DNNs to a small number, as the number of neurons in each layer is strongly correlated with execution time.

The Us-DMN architecture comprises four branches, which can be expanded depending on the number of modalities. In Figure 1, the top branch consists of an LSTM followed by a CNN. This branch is used for data with a sequential structure (in our case, the text). The other three branches consist of DNNs, which are used for emotions,

data extracted from the headset EEG, and other data (such as the subject of the debate and the participant's identifier). Each branch specializes in extracting latent characteristics for its respective modality. The penultimate and last layers are fully connected layers. The penultimate layer represents the user features vector, a latent representation of the user. The last layer corresponds to the classes. The user features vector is directly linked to the last layer and contains all the information needed for discrimination. The learned vector can also be used for other tasks, such as behavior prediction, as it provides information about the user's current general state, which can be useful for adaptation.

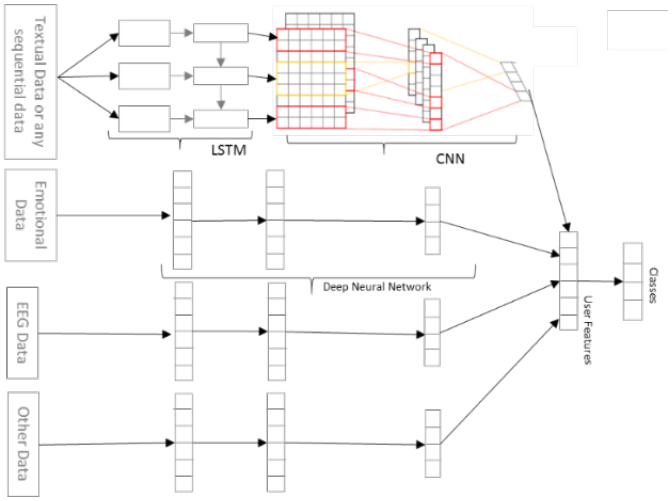

Figure 1: User-sensitive Deep multi-modal LSTM-CNN for user modeling

Experiments

Experiment Settings

Dataset : Seempad[1] is a public multimodal dataset of natural language arguments labeled with emotions and EEG data. Four people participated in 10 debates where everyone had to give their opinions. For each debate, a moderator gave instructions to the participants. For our experiment, w did not consider the arguments emitted by the moderators Thus, the corpus of data comprises 642 arguments (texts o at least one sentence), of which 127 have been labeled b the experts (attribution of the polarity of the opinion: Posi tive, Negative, Neutral). There are a total of 38764 rows i the dataset. Each row corresponds to the information (th seven emotions + Workload + Engagement) recorded ever second.

Parameters : Textual data (opinions) was represented us ing the publicly available GloVe [2] (Global Vectors for Wor Representation) vectors which have a dimensionality of 30(Words not present in the set of pre-trained words were initialized randomly. The number of DNN layers may vary depending on the context. For our application context, we have set the depth of the emotional data branch to 2 and the branches of the EEG and other data to 1. This is explained by the fact that the vector size of the EEG data is two and, therefore, already small enough for extracting hidden information and other data (participant and debate). As the CNN is a slightly modified version of the one proposed by Kim (Kim 2014), some parameters were chosen based on their results. For regularization, we employ dropout on the last layer of the CNN. Dropout prevents overfitting by randomly dropping out a proportion of the hidden units during forward back propagation (Srivastava et al. 2014). The dropout rate was set to .5, the mini-batch size is a size of 42, and we used 200 feature maps for each filter. Training is done through stochastic gradient descent with AdamOptimizer over randomly shuffled mini-batches. We used small filters size (3,4,5,6) because our dataset does not contain very long sentences. Since our 'best' size filter (Zhang and Wallace 2015) was 3, to determine the length and the size of the three other filters, we have just selected regions sizes near the best size. The size of the user latent vector features was fixed to 6. We trained the model on 65% and 85% of the labeled data, but the reported results as those from the second training.

[1]https://project.inria.fr/seempad/datasets/

[2]https://nlp.stanford.edu/projects/glove/

Results and Discussion

We compared our model to the widely used traditional methods in the same context (detecting the polarity of opinions): LSTM and CNN. We have four different versions of the two algorithms we have compared to our solution Figure 2. The first version is a simple one-layer LSTM (with the same architecture as the one used in our solution) with 200 hidden layers that take the GloVe representation of the arguments as input. The second version is the CNN (with the same architecture as the one used in our solution) which also takes the GloVe representation of arguments as input. The third version is the previous CNN, where the user data and extracted features are concatenated in the penultimate, resulting in CNN+User-data. The last version combines LSTM followed by CNN, which takes the GloVe vectors as input and is named CNN+LSTM.

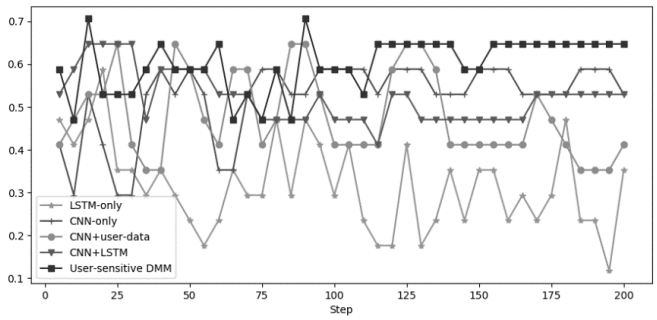

Figure 2: Accuracy of CNN, LSTM and our User-sensitive Deep multi-modal Network (Us-DMN).

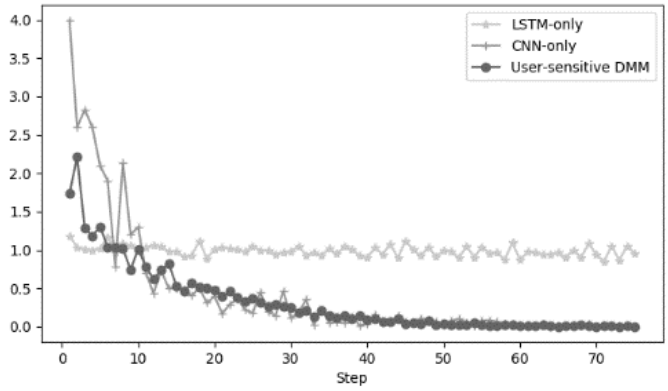

Figure 3: The evolution of the value of the loss on the training data.

While we had expected performance gains using pre-trained vectors with the LSTM or CNN model, we were surprised that they do not perform individually on our dataset. Results of our model against those methods are shown in Figure 2. At first glance, we can see that Us-DMN outperformed all other state-of-the-art techniques. We obtained 71% accuracy compared to 64%, 64%, 58%, and 47%, respectively, for the CNN+LSTM, the CNN+User-data, the CNN-only, and the LSTM-only models. One interesting thing we noticed in these results is that the CNN+LSTM and the CNN+User-data have the same performance, except that the first is more time-consuming than the second. Therefore, the prediction of user behaviors depends not only on user data but also on the architecture used to extract a representation of users. All the models were trained and tested only on the labeled data. We think it is the main reason the accuracy could not exceed 71%. However, our model can generalize better (still with little data) than the other methods presented.

Us-DMN is also very stable after many steps than the other techniques. Figure 3 shows the loss's evolution. The loss value of the LSTM-only version is very high compared to others. The evolution of the loss value of the CNN-only version is quite similar to that of the proposed model after about 20 steps. The loss correlates to the model's performance since it approximatively measures the distance between the expected and predicted values. The uses of the LSTM in our model did not affect his loss as we thought, which means adding user data definitively made the model converges better. It converges more quickly toward the optimal solution, requiring less training.

Our solution is a little slower in the training process than other techniques. This is mainly due to the complexity of the model architecture, which means more parameters to train. It may not be adapted to Time sensitive contexts, where the learning step is required every time new data is added to the system. The issue can be addressed by setting a fixed amount of new data or time intervals to renew the training process. However, such situations are infrequent, as we usually train the model before using it for prediction or classification.

Simply adding user data clearly improves the classification task, hence ensuring the relevance of the user features vector. Adding more information about the user is necessary to ensure the latent user vector is accurate. The more information we have, the better the model will be. Furthermore, we note that automatic behavior pattern detection heavily relies on user data, such as emotional states, and the architecture used to extract the user's latent representation.

Integrating DM-LTM in Various Applications

We have shown an excellent example of the practical application of Us-DMN, where ones try to detect the polarity of opinions automatically. The proposed solution can extract a latent representation of the user and use this rich representation to discriminate different behaviors. The model is adaptable to other contexts. For instance, we plan to integrate it into two learning environments. The first is a learning environment for doctors (a virtual operating room) to detect wrong actions against the good ones to provide adequate help to learners. The second is a first-person serious game called "LesDilemmes" , which aims to assess and train the social reasoning skills of the player. We plan to use the model to detect players' moral reasoning levels. The latent user features vector will represent the sociomoral facet of the player model. Necessary actions that need to be carried out to adapt the model to a new context include: 1) Determine the number of branches that vary according to the number of modalities; 2) Determine the modality (s) that contain sequential data (such as text) and send them as input to the dedicated branch (LSTM + CNN); 3) Define the depth of the DNNs for each modality according to how deep we want the model to look at the latent features; 4) Define the size of the latent user features representation vector that we want to extract and; 5) Define the size of the last layer (which depends on the number of classes or behaviors that we want to detect).

Conclusion

Us-DMN, a novel user-sensitive deep multi-modal model, was introduced in this study to enhance user modeling in platforms where adaptation is critical, including ITSs and serious games. The proposed model's applicability is wider than these domains and can be useful in any situation where user behavior management is necessary. Our research involved using Us-DMN to detect the polarity of people's opinions in various debates. The outcomes demonstrated that our model is superior to state-of-the-art architectures in a comparable task. It indicates that the penultimate layer of our architecture can extract latent features of the user to distinguish between different behaviors. The model's advantages over others are its adaptability and ability to operate in a multi-modal setting. Additionally, it can utilize the input data's sequential nature to enhance prediction and, as a result, the user's latent representation. The next steps in this research include evaluating Us-DMN in learning environments and improving the architecture to work with unlabeled data.

Acknowledgments

This work is supported by the Natural Science and Engineering Research Council of Canada Discovery Grant Program. http://www.nserc-crsng.gc.ca/index_eng.asp

References

Al-Dulaimi, A.; Zabihi, S.; Asif, A.; and Mohammadi, A. 2019. A multimodal and hybrid deep neural network model for remaining useful life estimation. *Computers in industry* 108:186–196.

Birk, M. V.; Toker, D.; Mandryk, R. L.; and Conati, C. 2015. Modeling motivation in a social network game using player-centric traits and personality traits. In *International Conference on User Modeling, Adaptation, and Personalization*, 18–30. Springer.

Covington, P.; Adams, J.; and Sargin, E. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, 191–198. ACM.

Dimitri, G. 2022. Survey on deep learning for multimodal integration: Applications, future perspectives and challenges. *Computers* 11(163).

Elkahky, A. M.; Song, Y.; and He, X. 2015. A multi-view deep learning approach for cross domain user modeling in recommendation systems. In *Proceedings of the 24th International Conference on World Wide Web*, 278–288. International World Wide Web Conferences Steering Committee.

Gadzicki, K.; Khamsehashari, R.; and Zetzsche, C. 2020. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd international conference on information fusion (FUSION)*, 1–6. IEEE.

Gunes, H., and Piccardi, M. 2007. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications* 30(4):1334–1345.

Huang, S.-C.; Pareek, A.; Zamanian, R.; Banerjee, I.; and Lungren, M. P. 2020. Multimodal fusion with deep neural networks for leveraging ct imaging and electronic health record: a case-study in pulmonary embolism detection. *Scientific reports* 10(1):1–9.

Jing, G.; Peng, L.; Zhikui, C.; and Jianing, Z. 2020. A survey on deep learning for multimodal data fusion. *Neural Computation* 32(5):829–864.

Kim, Y. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Liu, Q.; Huang, Z.; Yin, Y.; Chen, E.; Xiong, H.; Su, Y.; and Hu, G. 2019. Ekt: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering* 33(1):100–115.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, 689–696.

Nkambou, R.; Tato, A. A. N.; Brisson, J.; Kenfack, C.; Robert, S.; and Kissok, P. 2016. On the evaluation of the expert and the learner models of logic-muse tutoring system. In *Intelligent Tutoring Systems*, 506. Springer.

Piech, C.; Bassen, J.; Huang, J.; Ganguli, S.; Sahami, M.; Guibas, L. J.; and Sohl-Dickstein, J. 2015. Deep knowledge tracing. *Advances in neural information processing systems* 28.

Poria, S.; Cambria, E.; Hussain, A.; and Huang, G.-B. 2015. Towards an intelligent framework for multimodal affective data analysis. *Neural Networks* 63:104–116.

Ramesh, A.; Goldwasser, D.; Huang, B.; Daumé III, H.; and Getoor, L. 2013. Modeling learner engagement in moocs using probabilistic soft logic. In *NIPS Workshop on Data Driven Education*, volume 21, 62.

Sainath, T. N.; Vinyals, O.; Senior, A.; and Sak, H. 2015. Convolutional, long short-term memory, fully connected deep neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 4580–4584. IEEE.

Song, X.; Li, J.; Tang, Y.; Zhao, T.; Chen, Y.; and Guan, Z. 2021. Jkt: A joint graph convolutional network based deep knowledge tracing. *Information Sciences* 580:510–523.

Song, X.; Li, J.; Lei, Q.; Zhao, W.; Chen, Y.; and Mian, A. 2022. Bi-clkt: Bi-graph contrastive learning based knowledge tracing. *Knowledge-Based Systems* 241:108274.

Srivastava, N., and Salakhutdinov, R. R. 2012. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, 2222–2230.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15(1):1929–1958.

Tang, D.; Wei, F.; Yang, N.; Zhou, M.; Liu, T.; and Qin, B. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 1555–1565.

Tato, A.; Nkambou, R.; Brisson, J.; and Robert, S. 2017a. Predicting learner's deductive reasoning skills using a bayesian network. In *International Conference on Artificial Intelligence in Education*, 381–392. Springer.

Tato, A.; Nkambou, R.; Dufresne, A.; and Beauchamp, M. H. 2017b. Convolutional neural network for automatic detection of sociomoral reasoning level. In *Proceedings of the 10th International Conference on Educational Data Mining*, 284–289. International Educational Data Mining Society.

Wang, H.; Wang, G.; Li, G.; Peng, J.; and Liu, Y. 2016. Deep belief network based deterministic and probabilistic wind speed forecasting approach. *Applied Energy* 182:80–93.

Zeng, Z.; Tu, J.; Liu, M.; Huang, T. S.; Pianfetti, B.; Roth, D.; and Levinson, S. 2007. Audio-visual affect recognition. *IEEE Transactions on multimedia* 9(2):424–428.

Zhang, Y., and Wallace, B. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.

Zhang, C.; Fan, W.; Du, N.; and Yu, P. S. 2016. Mining user intentions from medical queries: A neural network based heterogeneous jointly modeling approach. In *Proceedings of the 25th International Conference on World Wide Web*, 1373–1384. International World Wide Web Conferences Steering Committee.

Zukerman, I., and Albrecht, D. W. 2001. Predictive statistical models for user modeling. *User Modeling and User-Adapted Interaction* 11(1-2):5–18.